1984-06

Draftsman displays: a graphical technique for exploratory data analysis

Johnson, Malcolm D., Jr.

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

DRAFTSMAN DISPLAYS, A GRAPHICAL TECHNIQUE
FOR EXPLORATORY DATA ANALYSIS

by

Malcolm D. Johnson Jr.

June 1984

Thesis Advisor: P.A.W. Lewis

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) Draftsman Displays, a Graphical Technique for Exploratory Data Analysis	5. TYPE OF REPORT & PERIOD COVERED Master's Thesis June 1984
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Malcolm D. Johnson Jr.	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93943	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93943	12. REPORT DATE June 1984
	13. NUMBER OF PAGES 102
14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)	15. SECURITY CLASS. (of this report)
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

GRAFSTAT, Multivariate data, Graphical analysis, Exploratory data analysis, Jittering.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This thesis develops and explores the graphical analysis of multivariate data sets through the use of a Draftsman technique of scatter plot displays. These plot displays are useful for determining associations and relationships between variables in order to promote an understanding of the characteristics of the data in exploratory and descriptive applications. General graphical enhancement techniques such as jittering and transformations are discussed and incorperated in the development of a

computer program which produces Draftsman displays. A technical description of the Draftsman computer program is presented, and user implementation procedures discussed. An analysis is conducted on two varied sets of data to demonstrate the versatility and utility of the Draftsman display technique for exploring data structures.

Approved for public release; distribution unlimited.

Draftsman Displays, a Graphical Technique for Exploratory Data Analysis

by

Malcolm D. Johnson Jr.
Captain, United States Army
B.S., United States Military Academy, 1976

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 1984

ABSTRACT

This thesis develops and explores the graphical analysis of multivariate data sets through the use of a Draftsman technique of scatter plot displays. These plot displays are useful for determining associations and relationships between variables in order to promote an understanding of the characteristics of the data in exploratory and descriptive applications. General graphical enhancement techniques such as jittering and transformations are discussed and incorporated in the development of a computer program which produces Draftsman displays. A technical description of the Draftsman computer program is presented, and user implementation procedures discussed. An analysis is conducted on two varied sets of data to demonstrate the versatility and utility of the Draftsman display technique for exploring data structures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. MOTIVATION

Recent advances in computer hardware and sofware capabilities have made available to a larger number of users powerful diagnostic and analytical tools for exploring data. These same advances however are responsible for a tremendous increase in the amount of data produced and available for analysis. Contrary to mathematical intuition, the availability of more data available does not always lead to greater precision in subsequent analysis. Often the increased amount of data confounds the analysis by overburdening our ability to process the information in a timely and understandable fashion.

Graphical displays are a method of visually portraying vast amounts of qualitative information. The primary benefit of graphical techniques is that the human eye-brain system has a powerful information processing capability. By maximizing our visual capability to process properly displayed data, we can rapidly summarize information, focus on salient features, discern abberations, and extract details of interest from a data set.

B. SCOPE

The purpose of a Draftsman display is to use the visual impact of an array of two dimensional scatter plots to analyze multivariate data. This can be accomplished by arranging an exhaustive series of plots consisting of all paired variables. This enables the analyst to observe the influence of each variable on every other variable in the data set.

The concept of using two dimensional scatter plots to display higher dimensional data structures is discussed in a text by Chambers, et al. [Ref. 1]. The ideas from that text served as the foundation for the development of an interactive computer program to construct Draftsman displays. The additional features of enhancing scatter plots such as the jittering of discrete data values and transforming variables can also be applied to this multidimensional display procedure. The full considerations that went into the program development as well as user implementation procedures is amplified in later chapters.

The purpose of this thesis is to integrate the graphical concepts of scatter plots, jittering, and transforming variables into a Draftsman display. Although written in A Programming Language (APL), little if any knowledge of this language is required to successfully utilize the program.

This thesis has been written in three major segments in order to appeal to the widest audience possible. The first segment, composed of chapters II and III deals with the general concepts of graphical methods and user instructions required to invoke the Draftsman display program. The second segment, comprised of chapter IV and Appendix A, is aimed at those readers interested in the technical details and Draftsman program documentation. The final segment, found in chapters V and VI, contains a stepwise analysis of two varied forms of data to demonstrate potential applications of this procedure in exploratory data analysis.

The graphs used in this paper were produced by an experimental APL package GRAFSTAT, which the Naval Postgraduate School is using under a test agreement with the IBM Watson Research Center, Yorktown Heights, New York. We are grateful to Dr. P.D. Welch and Dr. P. Heidelberger for making GRAFSTAT available to us.

II. GRAPHICAL TECHNIQUES

A. DATA DISPLAYS

A variety of graphical methods such as box plots, histograms, stem and leaf displays and scatter plots are available to explore relationships which may exist between variables of a data set. The scatter plot is perhaps one of the most powerful graphical methods for displaying bivariate data. The foremost feature of the scatter plot is that all of the data of interest is readily displayed for visual interpretation. In addition, the simplicity of construction, compactness of the display, and adaptability to other graphical enhancement techniques, contribute to the power of this display.

In contrast, numerical summaries may reflect correlation but tell little about clustering, patterns, or other relationships which might be present. This is particularly true of larger data sets consisting of more than twenty observations and more than two variables. In these larger data sets, the sheer volume of data points to be compared makes interpretations a tedious and time consuming process.

Figure 2.1 is a scatter plot of weight versus engine displacement for 106 different models of cars produced during 1983 [Ref. 2 :pp.320-356]. A numerical summary might readily impart the fact that an increase in car weight is associated with an increase in engine size. The scatter plot however rapidly makes apparent some other interesting features. We can see that the observations consist of two distinct groupings. For vehicles under 3,000 lbs there is a strong positive linear dependency between weight and displacement. For the heavier vehicles over 3,000 lbs,

increasing weight still tends to be correlated with large displacements, though more dispersed in form. A numerical summary would not so easily reveal these relationships.

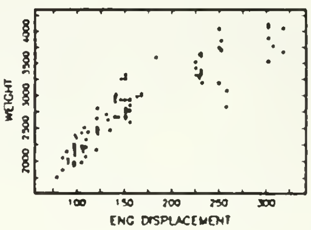


Figure 2.1 Basic Scatter Plot.

For many applications, our interest may extend beyond bivariate data sets to larger multidimensional sets. As in the bivariate case, scatter plots may also be used to graphically display multivariate data sets. An exhaustive series of plots consisting of all paired variables performs a similiar function as the single scatter plot does for bivariate data. By properly aligning the plots so that a commonality of axis exists between every plot and the adjacent plots, we can not only observe the relationships within a specific plot but may also follow particular obervations or groups of observations through the succesive plots to analyse the influence of other variables. This particular technique of arranging the scatter plots is similiar to a draftsman drawing of a three dimensional object and hence is termed a Draftsman display. [Ref. 1 :p. 136]

The three dimensional draftsman display shown in figure 2.2 consists of the variables of weight, turning radius, and engine displacement for 1983 model cars. The first row shows the paired plots of weight versus turning radius and engine displacement. The second row is turning radius versus weight and engine displacement. The bottom row of

plots consist of engine displacement versus weight and turning radius. This arrangement of the plots, while somewhat redundant, allows the viewer to scan across rows or down columns of plots, thereby matching up points that correspond to the same observations in different plots.

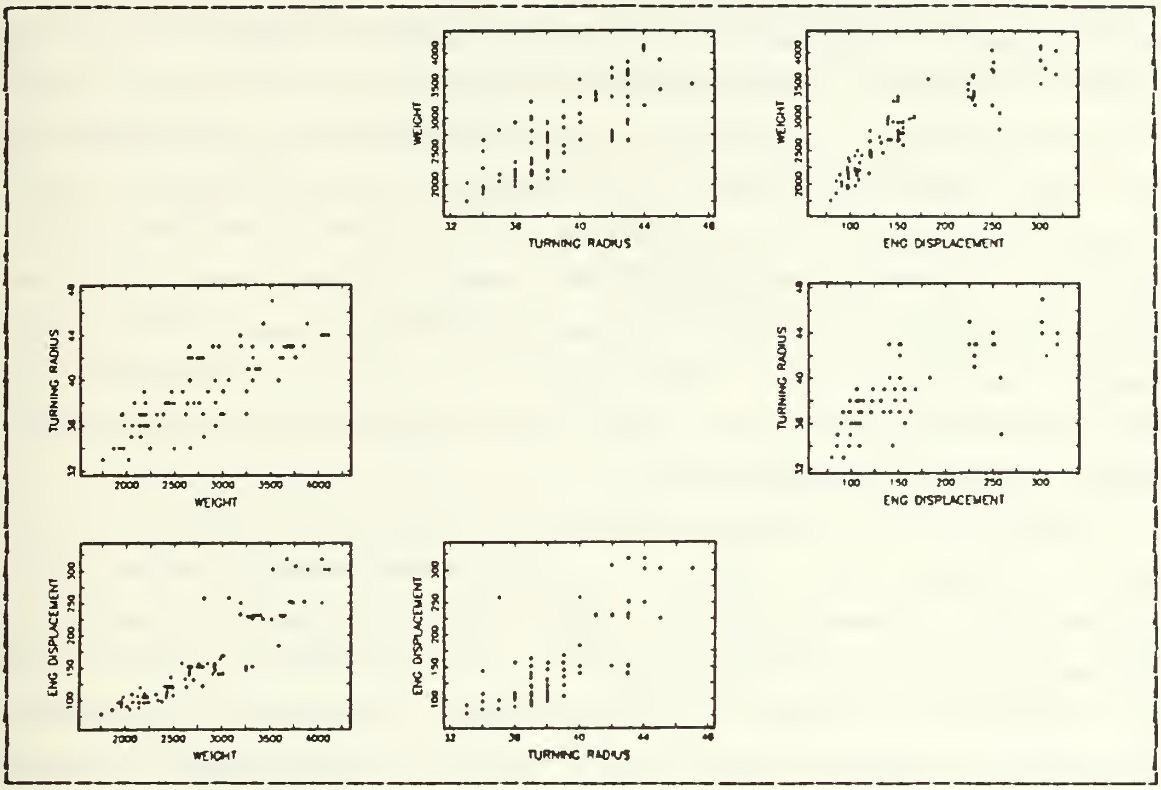


Figure 2.2 Three Dimensional Draftsman Display.

Observing the bottom row of plots in figure 2.2 , we can track three distinct groupings of points through all the paired plots. These three groups correspond to the small, medium, and large size categories of vehicles. A quick look at the associations exhibited in this display indicates also that engine displacement has a tighter relationship with weight than it does to turning radius. Other relationships are also evident and are presented in greater detail in the analysis presented in Chapter V.

B. JITTERING OF VARIABLES

In certain circumstances, a scatterplot may itself be visually. deceiving due to the overlapping of data points within the plot. This situation may be particularly prevalent when one or both of the plotted variables have a limited range of discrete values. In order to alleviate the overlapping and enhance the actual relationships that exist, a small amount of random noise may be added to one or both of the variables to "jitter" their horizontal and vertical locations within the plot. The amount of random noise added or subtracted from the original data values must be sufficient to prevent overlapping but small enough so that the original data values can be recovered by rounding to the nearest whole number. Typically the random noise added is two to five percent of the total range of the variable values. [Ref. 1 :pp.106-107]

The visual difference resulting from jittering can be seen in figure 2.3, where the maintenance records for 1981 versus 1982 was plotted for 106 automobile models. Maintenance is a category variable with values of 0, 1, ..., 5. Clearly a problem of overlapping exist in the basic scatter plot seen on the left. The jittered version on the right is a more accurate picture of the distribution and clustering prevalent in the data.

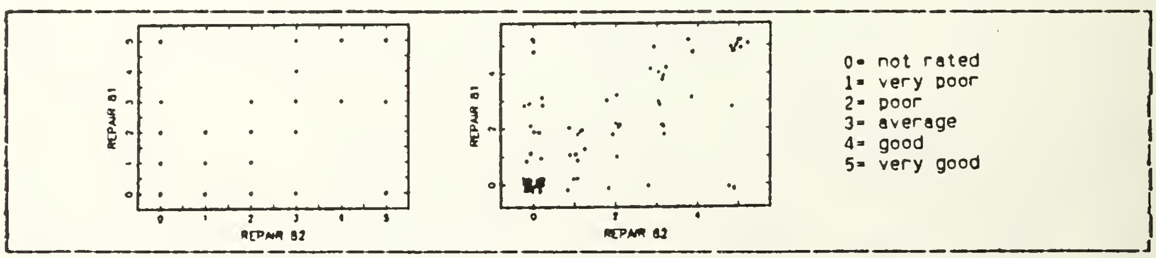


Figure 2.3 Unjittered and Jittered Plots.

C. TRANSFORMATION OF VARIABLES

The primary purpose of employing transformations is to linearize and simplify the observed relationship between the variables plotted. In many instances the plots may be further enhanced through the use of transformations in order to achieve a simpler and more understandable picture suitable for visual comparisons and exploration. Hoaglin [Ref. 3 :p. 104], proposes the following pertinent reasons for employing transformations.

1. Facilitate interpretation in a natural way.
2. Promote symmetry in a batch.
3. Promote stable spread in several batches.
4. Promote straightline relationships between the variables.
5. Simplify the structure of a two way or higher dimensional data structure so that a simple additive model can assist in the understanding of the characteristics of the data.

A key factor of transforming the variables is that if the correct transformation is applied, the resulting scatter plot will appear more linear in form. This in turn visually enhances recognition, detection of deviations and outliers, and assist in observing relationships or patterns.

As previously discussed, the basic scatter plot of weight and engine displacement is divided into two distinct groups of points as seen in the left plot of figure 2.4 . While the lower group appears fairly linear, the upper group is more dispersed and curved in shape. The plot on the right shows the effect of applying a log transform to the engine displacement values. The resulting plot of the transformed data becomes more linear over the entire range of engine displacement values (see figure 2.4).

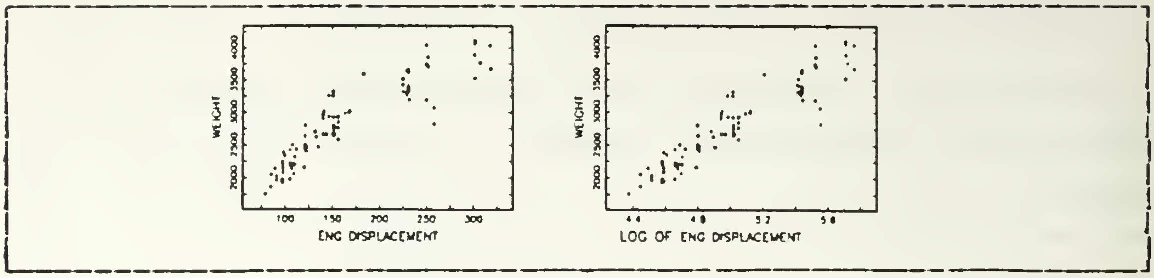


Figure 2.4 An Example of Transformation.

A note of caution is appropriate in determining when to use transformations in the Draftsman display. Since transformations result in a change to the displayed values and scale, care must be taken to avoid confusion during subsequent analysis. We should insure that the benefits of describing the data with a transformation is greater than the loss of simplicity incurred through its use.

III. DRAFTSMAN USER INSTRUCTIONS

A. GENERAL GUIDANCE

The Draftsman program was written in APL and is designed to be used in conjunction with the experimental IBM graphics software GRAFSTAT. The Draftsman program is interactive and requires little knowledge of APL to use. The APL versed user can easily modify the basic program and called subroutines for more specialized forms of analysis.

The graphical software which generates the Draftsman displays requires the use of either the IBM 3277GA or 3278/79 graphic display terminals [Ref. 4]. Normally these terminals are available as public facilities with special accounts and passwords. Once logged on to one of these terminals the user may link back to their own account and copy any of their own files as desired. This is useful in retrieving data files which the user wishes to analyse with a Draftsman display. [Ref. 5]

B. USER REQUIREMENTS

This section will provide a brief overview of the user inputs required to generate a Draftsman display of a data set. An explicit step-by-step description of all input requirements is found at the conclusion of this chapter in figures 3.2 through 3.6 .

Since the Draftsman program is written in APL, the user will have to enter the APL sub-environment in order to gain access to the graphics programs. Once in the APL environment, the APL characters set is invoked by keying the APL ON key. These APL characters are found in red and supercede the normal keys. It is recommended that the first time user

take a few minutes to familiarize themselves with the locations of these characters.

The APL environment will allow the user to copy and retrieve both the GRAFSTAT and Draftsman programs as shown in figure 3.2 . Once these programs are in the workspace the basic set up procedure is complete and the user is ready to actually initiate the Draftsman program to produce a display. The Draftsman program is initiated by typing DRAFTSMAN followed by return. The program will respond with a series of terminal queries requesting the various input parameters required in the display. Each query is generated based upon the user response to the previous query. The general program schematic and input requirements is outlined in figure 3.1 .

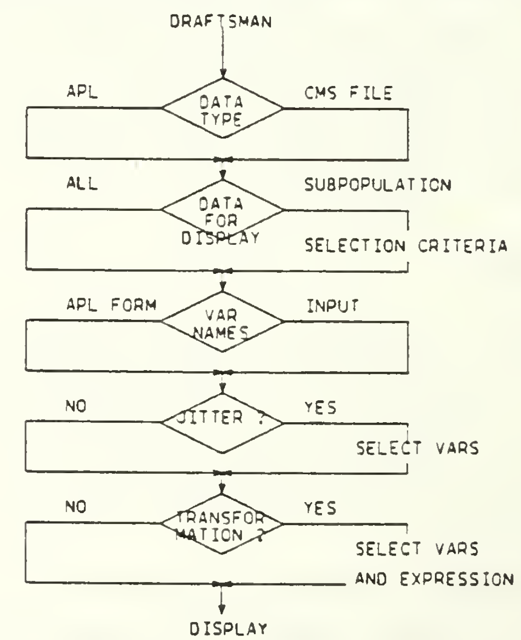


Figure 3.1 Draftsman Program Schematic.

The first option presented is that of inputing a data set (figure 3.3). Data which has been previously copied from another APL workspace may be entered by variable name. Data which is located on a CMS file can be automatically read into the workspace. Data in a CMS file can contain only numeric characters. A mixture of numeric and alphabetic characters will result in the data not being read in correctly. A crucial requirement is that regardless of how the data is entered it must be in two dimensional array form (rows and columns). The columns of the data correspond to the variables, the rows to the different observations on each variable.

Once the basic data has been entered the user is presented with an option to have either all of the data or only a subsample of the data appear in the display. This allows Draftsman displays to be produced on either all the data, specified variables, a subpopulation of a variable, or any combination thereof (figure 3.4).

Based on the data selected, an option will be presented to enter the appropriate names of the variables which will appear in the display (figure 3.5). These names are the labels which will appear on the axis of the plots. The variable names can be entered as a previously generated APL two dimensional array of characters. If this method of input is selected, each row of the array must contain the name of a variable in the same order as the variable is located in the data structure. The variable names may also be entered directly in response to a sequential series of queries. Once the variable names are entered, the minimum input requirements needed to produce a Draftsman display have been completed. The remainder of the queries pertain to display enhancements which may be invoked if desired.

The first enhancement option is that of jittering (figure 3.6). An input of 0 will result in no jittering of

the data. If jittering is desired, the user will be queried as to which variables. The results of jittering appear only in the Draftsman display and do not permanently alter any of the values in the original data set.

The second enhancement option available is transformation of variables (figure 3.6). Here again, a response of 0 will result in no transformations occuring. If one or more transformations are desired the user is prompted for the variables and APL expression for the transformation. A summary for some of the more common transformations with examples is illustrated in Table I .

TABLE I

Sample APL Transformations

TRANSFORM	MATH FORM	APL EXPRESSION
LOG	LN X	⍟X
LINEAR	$Ax+B$	B+A×X
CUBIC	X^3	X*3
CUBE ROOT	$X^{-1/3}$	X*(−1/3)
SQUARE	X^2	X*2
SQUARE ROOT	$X^{-1/2}$	X*(−1/2)

The Draftsman program will begin to display the component scatter plots on the graphics screen. The entire display is generated in segments of five variables. At the end of each segment an option is offered for the user to quit, continue, or to make a hardcopy and continue.

ENTRIES	TERMINAL RESPONSE
APL	VS•APL 4.0
	CLEAR WS
)LOAD 20 GRAFST3	SAVED 11:15:01 04/08/83
	.
	.
	.
	FOR MORE INFORMATION TYPE: DESCRIBE
)COPY DRAFTSMAN	SAVED 21:23:57 04/10/84

NOTE: GRAFST3 is one version of the experimental APL graphics program. If this procedure does not work, contact Professor P.A.W. Lewis (x2283,Root 269 Naval Postgraduate School) for more information.

Figure 3.2 Accessing the Graphics Programs.

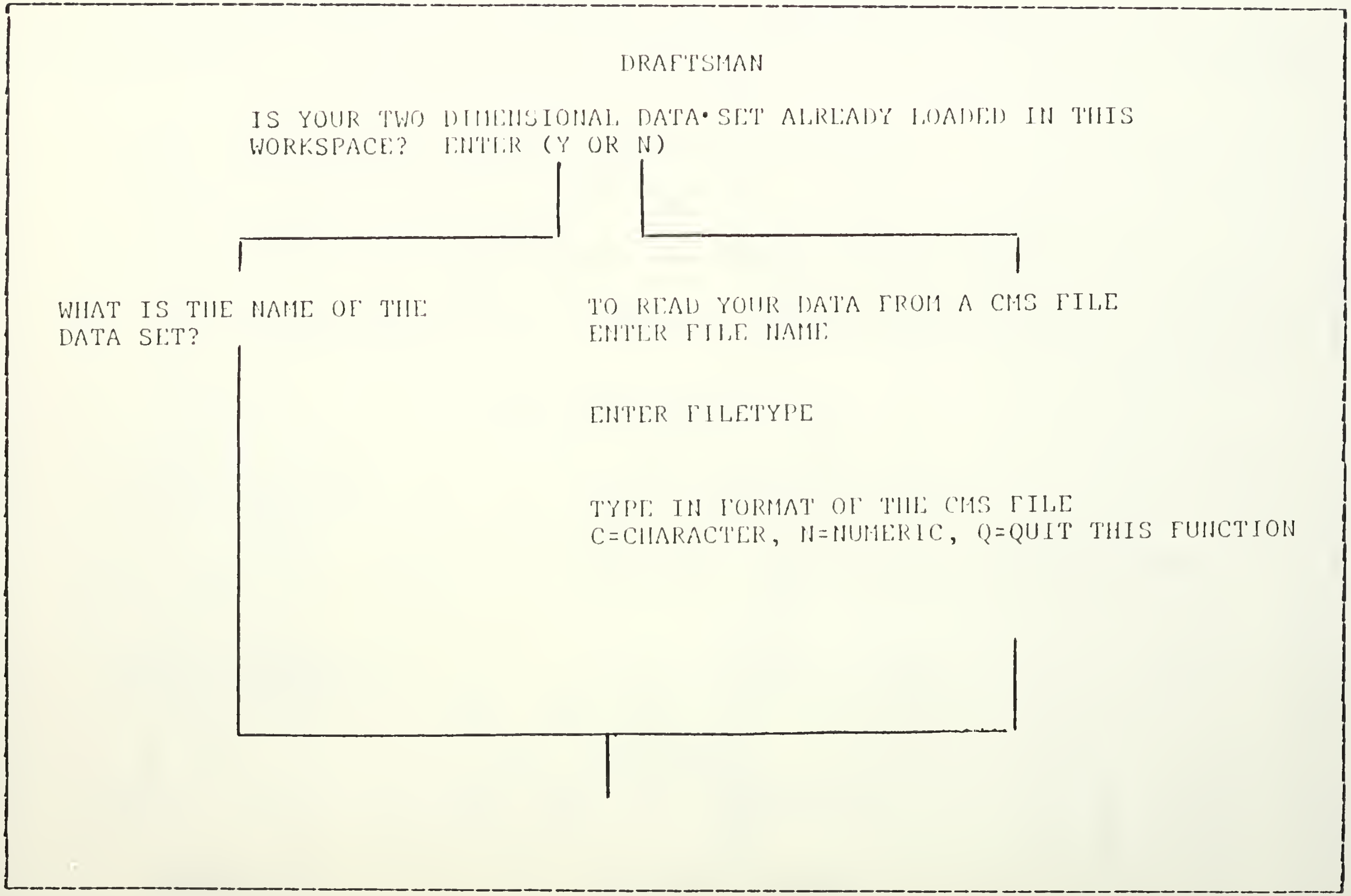


Figure 3.3 Data Input Options.

DO YOU DESIRE ALL OF THIS DATA PRESENTED IN THE DRAFTSMAN DISPLAY OR
JUST A SUBSAMPLE OF IT? ENTER (SUB OR ALL)

ENTER AS A VECTOR THE VARIABLES (COLUMN LOCATION)
OF THE DATA YOU WANT DISPLAYED

DO YOU DESIRE A SUBPOPULATION GROUP OF ANY
ONE VARIABLE? ENTER (Y OR N)

WHAT VARIABLE (COLUMN) IN THE ORIGINAL
DATA IS THE SUGROUP A MEMBER?

ENTER AS A VECTOR THE VALUES OF THIS VARIABLE

THE SUBDATA DESIRED IS A GLOBAL VARIABLE CALLED
DATA AND HAS A SHAPE OF __ __.

Figure 3.4 Data Subsampling Options.

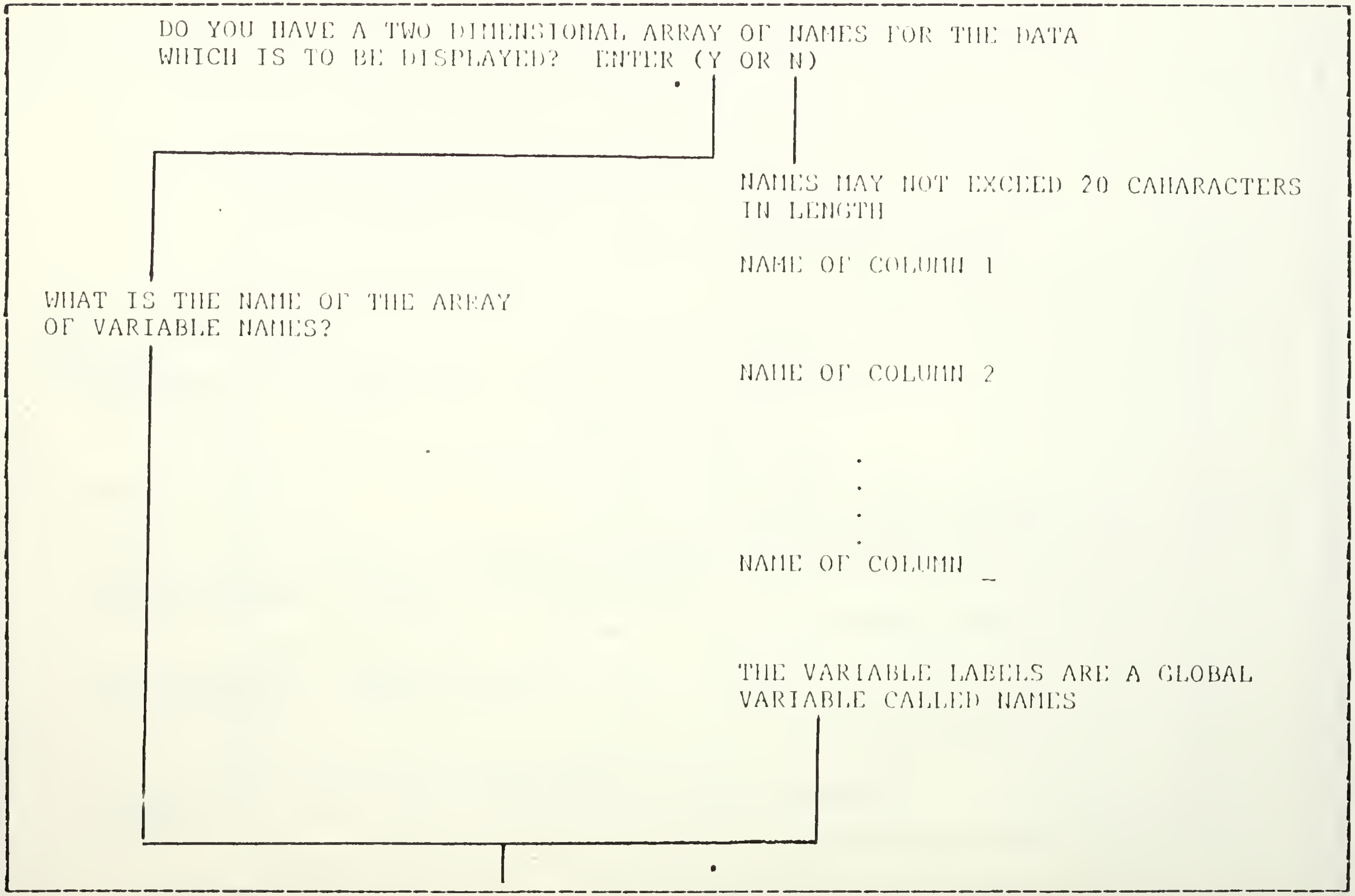


Figure 3.5 Variable Labeling Options.

HOW MANY VARIABLES DO YOU DESIRE JITTERED?

WHAT ARE THE VARIABLES (COLUMNS)
TO BE JITTERED?

HOW MANY VARIABLES DO YOU WANT TO HAVE TRANSFORMED?

WHAT ARE THE VARS (COLUMNS) TO BE TRANSFORMED?

USING X AS THE VAR, INPUT THE APL EXPRESSION
FOR THE TRANSFORMATION DESIRED ON COLUMN _

.
.
.

USING X AS THE VAR, INPUT THE APL EXPRESSION
FOR THE TRANSFORMATION DESIRED ON COLUMN _

Figure 3.6 Display Enhancement Options.

IV. DRAFTSMAN TECHNICAL IMPLEMENTATION

A. BASIC DRAFTSMAN ROUTINE

The Draftsman program was written in APL, which is an array processing language. The use of APL enables the Draftsman program to call and implement a variety of plotting functions available in the IBM GRAFSTAT graphical software package. The GRAFSTAT software is an experimental graphics program currently under development by IBM. It is presently available at the Naval Postgraduate School for testing and evaluation purposes. [Ref. 4]

A secondary benefit derived in using APL is the inherent user efficiency characteristics in terms of the large number of mathematical operations executable directly as keyboard entries. This approach is ideal for exploring data structures and features of interest. [Ref. 6]

The foundation of the Draftsman program revolves around the graphical plotting features of GRAFSTAT, and in particular the scatter plot option. This option requires the user to input the two variables of interest, size and location of the plots, as well as any headings desired. Figure 4.1 is an example of the scatter plot input screen, called The alphanumeric screen (ANS) in GRAFSTAT.

Correct alignment of the series of scatter plots so that a commonality of axis between adjacent plots exist involves an automatic reiterative calling of the scatter plot graphics function. Each input parameter of the basic scatter plot function is assigned as a local program variable. Based upon the input data structure, the Draftsman program sytematically selects the variables to be plotted, appropriately labels each axis, and determines the correct

```
                              PLOT FUNCTION

 X VARIABLE(S)                : X
 Y VARIABLES(S)               : Y
 TYPE(S) OF PLOT              : 0
 TYPE(S) OF LINE              : 1
 TYPE(S) OF SYMBOL            : .
 PLOT HEADER (IN QUOTES)      : ''

 SCREEN HEADER (IN QUOTES)    : ''

 X-AXIS LABEL (IN QUOTES)     : XAXIS
 Y-AXIS LABEL (IN QUOTES)     : YAXIS
 POSITION                     : POSN
 SCALE X-AXIS : LIN LX TX                  SCALE Y-AXIS : LIN LY TY
 PARTIAL PLOT                 : 1 1 1
 AXES AND GRID CONTROL        : 10 11 0 0


 ENTER=GO        PF:1=HELP  2=VIEW GRAPHICS(3279)  3=RETURN   4=WRITE ON SCREEN
 CLEAR=DEFAULT      5=LAST RESPONSES    6=ERASE      7=AXIS CONTROL     8=LEGEND
      RESPONSES     9=OUTPUT  10=STORE/RETRIEVE   11=INTO APL  12=SCREEN DISPLAY
```

Figure 4.1 GRAFSTAT Scatter Plot Function Screen.

location in which each of the plots will appear for display. This methodology produces the entire array of plots while eliminating the need for reiterative inputs by the user for each plot. The output is displayed as a row of scatter plots for each variable in the data set.

For data structures consisting of five or less variables, the Draftsman program display will fit on a single page. The plotting of five variables per page was selected to balance space limitations against the need for sufficient clarity of detail within the plots. To accomodate more than five variables on a single page would require smaller plots while reducing the visual usefullness of the display and making comparisons inconvenient. Less than five variables per page results in the excessive use of costly graphic reproducing paper. For data sets exceeding five variables, the Draftsman display is generated in segments which when reproduced may be pasted together to form a completed display.

The segmental method of producing Draftsman displays enables the user to display data sets consisting of more than five variables. The display procedure is limited only by the workspace capacity of the user computing facility. The number of segments that will result in a Draftsman display can be calculated by squaring the number of variables in the data set and dividing by 25. In practice, a display of more than 15 variables becomes somewhat unwieldy and may negate the benefits of using a Draftsman methodology.

B. TECHNICAL DETAILS OF INPUT REQUIREMENTS

A two dimensional array of data and a two dimensional array of the data variable names are the minimum input parameters required to generate a Draftsman display. These parameters are inputted as prompted by the routine ADMIN.

Data may be input directly as an APL variable or retrieved from a CMS file located on the user's disk. File reading is accomplished by CMSREAD [Ref. 6], a library routine which has been pre-copied into the Draftsman workspace.

A program entitled SUB was written to assist in the restructuring of data sets into more convenient formats. An initial analysis of the basic Draftsman display may reveal certain variables or sections of data points which warrant closer scrutiny. The SUB program allows the user to select variables from the original data set as well as subsamples of a variable in order to create a new data set entitled DATA. DATA becomes the global variable that is actually displayed. The APL program SUB which implements this procedure is found in Appendix A.

The matrix of variable names is either input directly as an APL two dimensional array or is generated by the routine

LABELS. The rows of the matrix correspond to each of the variables in the data set which is to be displayed. When generated by LABELS, each variable name entered containing less than 20 characters is padded with blank spaces. If more than twenty characters are entered, only the first twenty characters will appear on the display. This assures that the entire array when passed to succeeding routines is a valid rectangular character array. The LABELS routine which implements this procedure is found in Appendix A.

C. ENHANCEMENT ROUTINES

1. Jitter Routine

As discussed in Chapter II, overlapping of plotted values may be misleading and inadequately portray the visual relationships exhibited in the data structure. The solution is to jitter or add random noise to one or both variables to be plotted. This technique is presented as an enhancement option to the user and requires only an identification of the variables upon which jittering will be performed.

The jittering of variable points within the Draftsman program is accomplished through a method discussed by Chambers [Ref. 1 :pp. 106-107]. We let Ui, for i=1 to n (the number of observations), be n equally spaced values from -1 to +1 in random order. The original variable values are thus reexpressed in jittered form Ji,

$$J_i = X_i + \theta_x U_i \qquad \text{(eqn 4.1)}$$

where θx is .05 times the range of the variable data values. This method results in a fractional shift of the data values along the same axis in which the variable is plotted.

The small shifting of plotted points is sufficient to negate the effects of overlaps while preventing any serious corruption of the plotted data. It shows the multiplicity of points at each actual coordinate. The original data values can be recovered by rounding to the nearest integer. Internally the Draftsman program uses the original data to create local variables which are jittered and then plotted. This enables the user to always maintain the data in original form.

2. Transformation Routine

The potential for transforming variables was written as a user option to further enhance the basic Draftsman display. This routine maximizes the characteristics of the APL primitive functions as well as parallel array processing. The program requires the identification of the variables desired to be transformed and the appropriate APL expressions for each transformation. The parallel processing capability transforms each variable set in one operation. As in the jitter routine, the transformation routine transforms only local variables for plotting and leaves the original data structure intact.

V. AN ANALYSIS OF AUTOMOBILE DATA

A. INTRODUCTION

An analysis is presented of data consisting of selected characteristics of automobiles manufactured during 1983 and tested by Consumers Union. The primary purpose of this chapter is to demonstrate an application of Draftsman displays in exploratory data analysis. The analysis initially explores the general descriptive qualities of the characteristics of automobiles using the basic Draftsman display procedures. Subsequent analysis focuses on observed variables of interest as developed through the enhancement features of Draftsman.

B. THE AUTOMOBILE DATA

The data was initially formatted as a two dimensional array consisting of 106 rows and 14 columns. Each row of the data matrix corresponds to one of the 106 different models of automobiles as tested by Consumer Union [Ref. 2 :pp.320-356]. The columns contain various characteristics for each of the automobiles. These fourteen variables comprise the three general categories of price, performance, and size. The price category consists of the suggested retail price of the basic automobile without additional options. The performance variables include fuel efficiency (city and highway), turning diameter, gear ratio, and vehicle repair records for the two preceeding years. The size variables consist of length, weight, headroom, rear seating space, trunk size, and engine displacement. A general variable, automobile, corresponds to each of the specific models upon which the data is based. A summarized description of the data is shown in Table II for reference.

TABLE II

Automobile Data Characteristics

VARIABLE	UNITS	REMARKS
Automobile		1 to 48; small cars 49 to 98 mid cars 99 to 106; large cars
Price	$1000	
MPG City	Miles per gallon	EPA rated
MPG Highway		
Repair 81		0= not rated 1= very poor 2= poor 3= average 4= good 5= very good
Repair 82.		
Headroom	Inches	
Rear seat space	Inches	
Trunk size	Cubic feet	
Weight	pounds	
Length	inches	
Turning radius	Feet	
Engine displacement	Cubic inches	
Gear ratio		

C. PRELIMINARY ANALYSIS

1. General

The general Draftsman display of variables was generated as discussed in chapter II. A reduced version of the basic displays is seen in figures 5.1 through 5.4 . The actual Draftsman displays used for analysis may be found in Appendix B. For convenience and clarity, individual scatter plots from the displays will be included within applicable sections of the text.

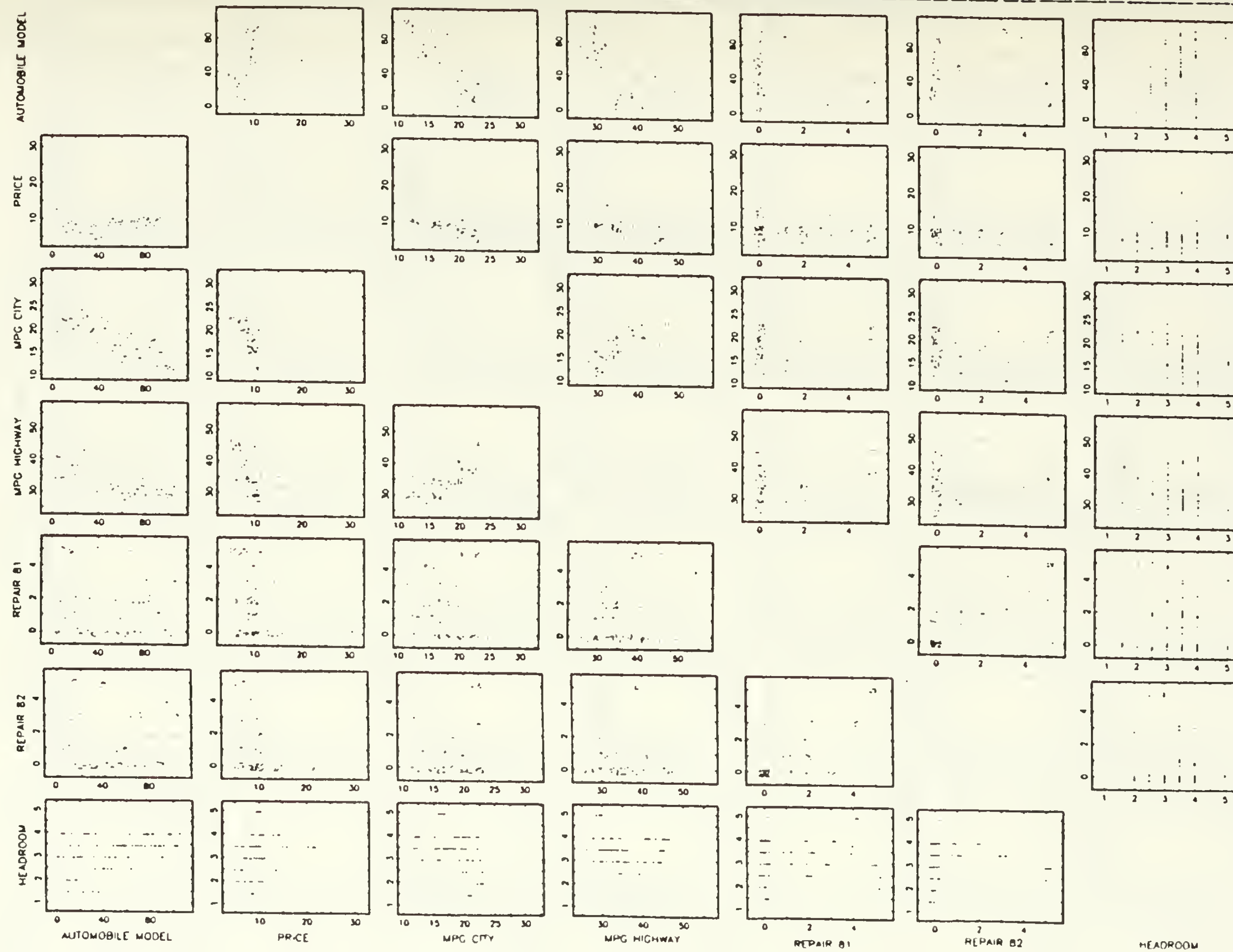


Figure 5.1 Segment 1 of Automobile Data.

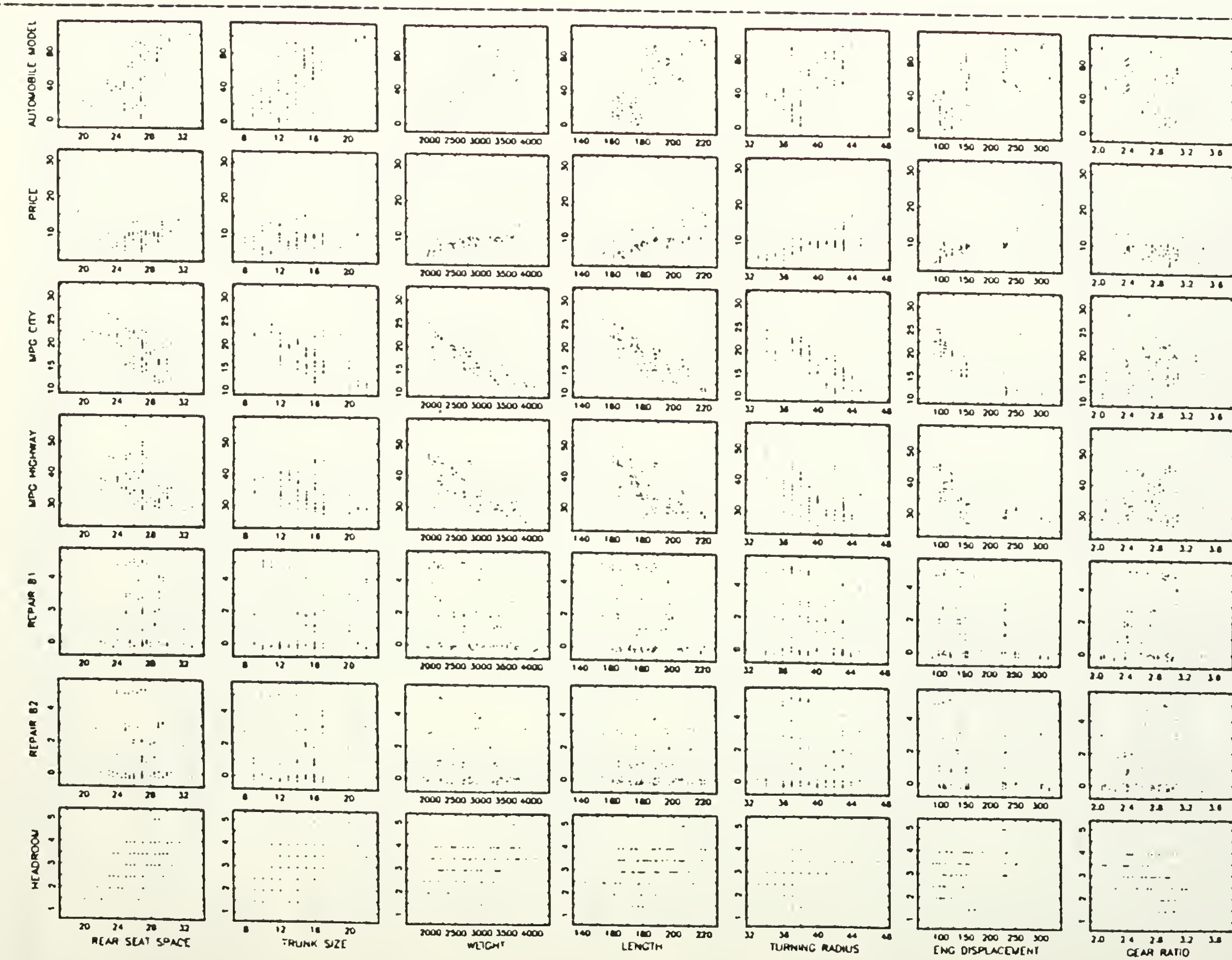


Figure 5.2 Segment 2 of Automobile Data.

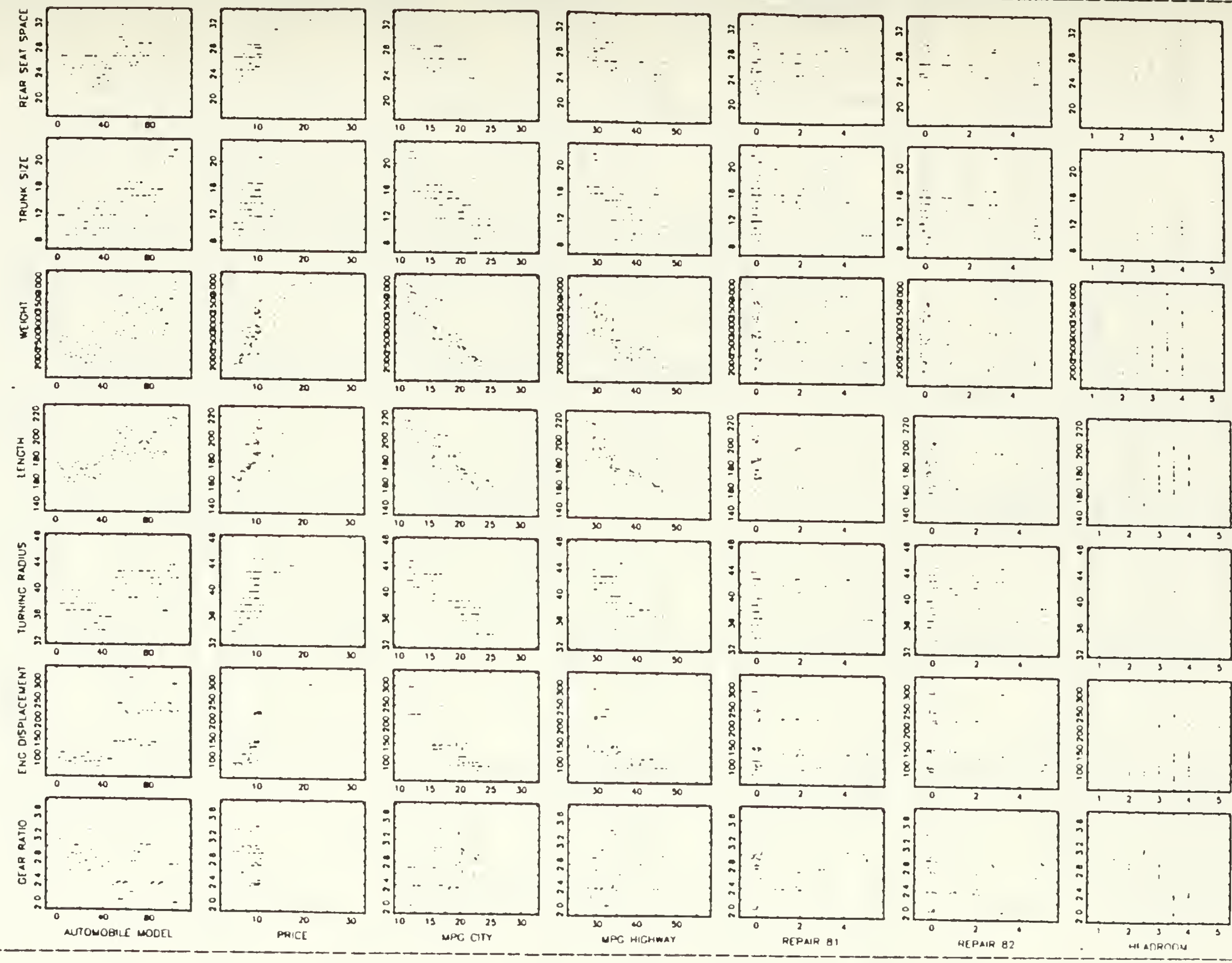


Figure 5.3 Segment 3 of Automobile Data.

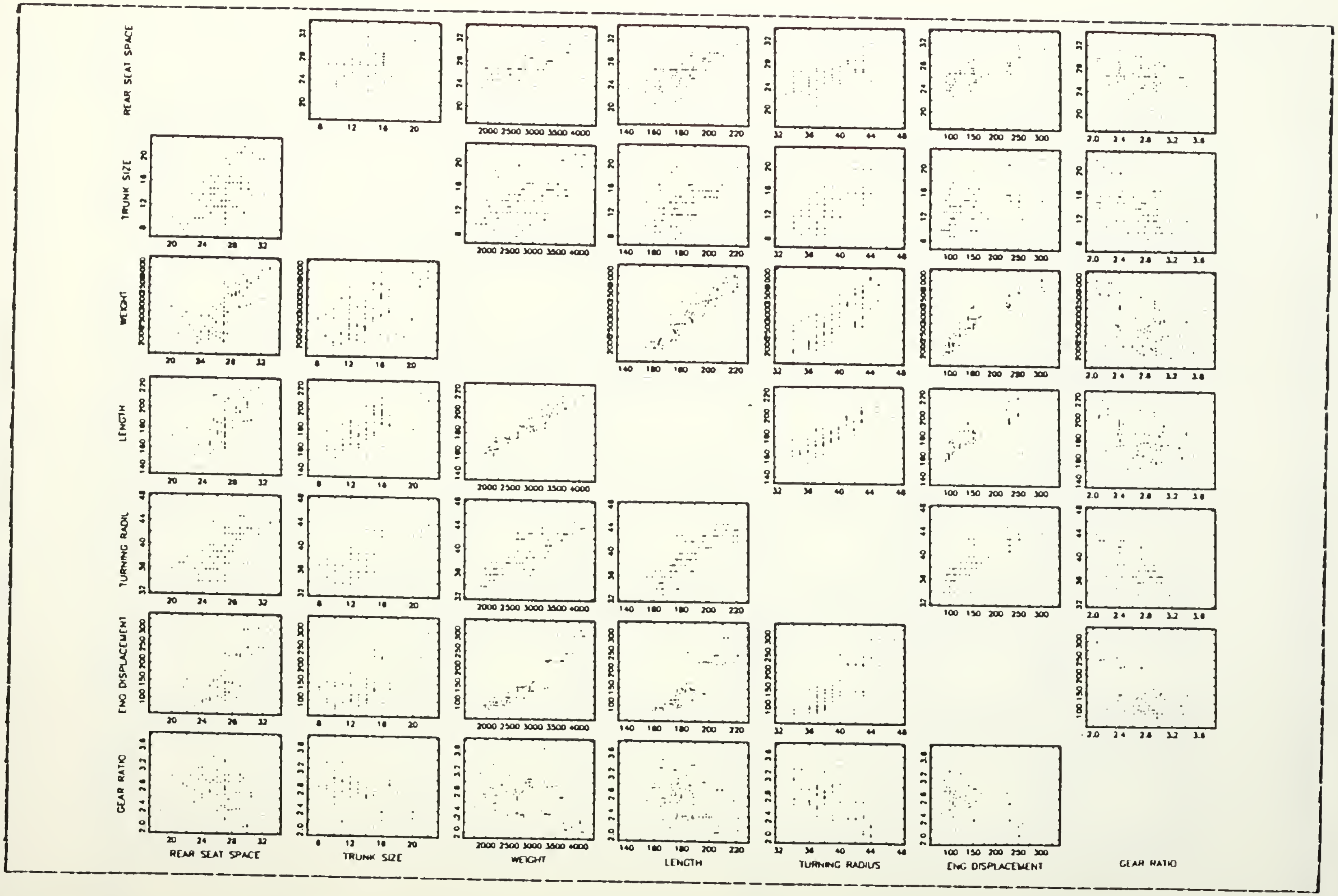


Figure 5.4 Segment 4 of Automobile Data.

The general characteristics of automobiles are likely to be familiar to most readers. Intuitively we can perhaps surmise many of the relationships of the data structure without even looking at it. This familiarity however will enable us to concentrate more on the features of the Draftsman program in exploring the data. Additionally, we may confirm intuitive knowledge or perhaps change some of our perspectives based upon the analysis.

2. Characteristics of Price

Focusing on price as relating to the other parameters, the first visual message imparted in figure 5.5, is that price bands deliniate the major categories of automobiles. Generally the small sized cars are grouped at under $10,000 while midsize models are rather tightly grouped between $7,500 and $12,000. If we concentrate on deviations from this pattern, the outliers reveal an interesting feature. From each major size category to the next there is a substancial increase in the number of outliers within the categories. These outliers are predominately luxury models within their respective categories.

When price and weight are compared, a gentle upward sloping trend dominates, denoting that price and weight are positively related, which is to be expected (figure 5.5). This relationship levels off at about $10,000. A very obvious branch from the main trunk of observations shows price increasing relative to weight at a greater rate. The presence of this branch suggested additional research to determine if a significant parameter was missing from the data. The research revealed this uppermost branch consisted of luxury style models, with all but one of foreign manufacture. The majority of the outliers contained between the two branches are the luxury models of American origin. We might conclude that weight is generally associated with an

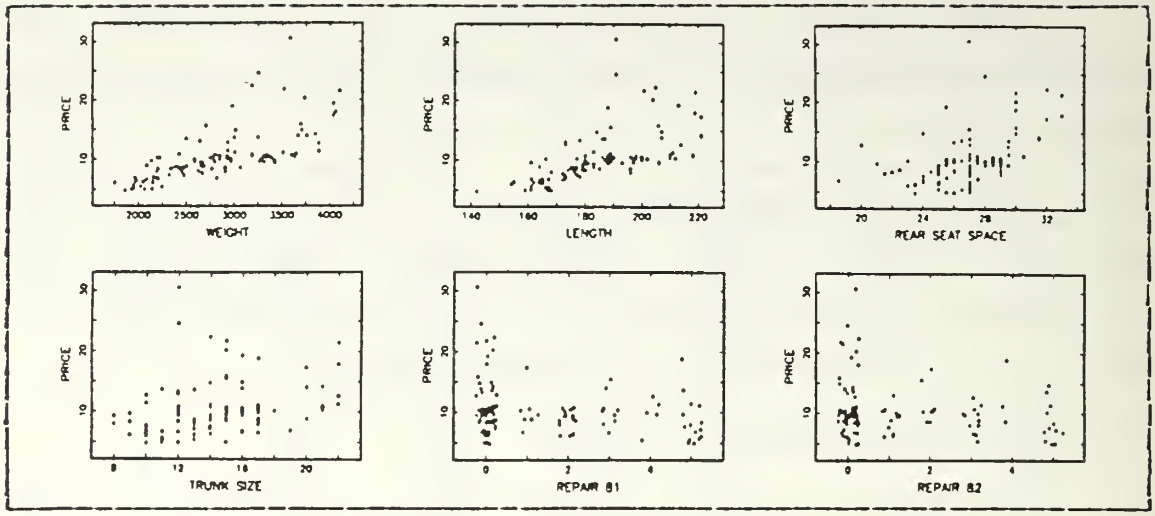


Figure 5.5 Characteristics of Price.

increase in price, with the foreign luxury models tending to be more expensive than American luxury models of comparable weight.

Similiar upward curving relationships as that observed in the scatter plot of price versus weight can be seen in some of the other plots in the price row in figure 5.5 . The plot of price and weight is most closely resembled by that of price and length. This however may be somewhat deceiving. A little thought might lead us to conclude that these similarities are more due to a relationship between length and weight. Consistent with this are the two plots containing the parameters of rear seating space and trunk size versus price. Although they loosely resemble the pattern of the price versus weight plot, we should suspect that they are influenced more by the overall dimension of automobile length. These plots demonstrate the care that must be taken in the analysis of single scatter plots. We must be cautious since each scatter plot in the array denotes only the isolated relationship of two variables and may not necessarily indicate a causal relationship.

In order to eliminate the overlapping of the discrete valued maintenance ratings, jittering was required. If we look at price with respect to the two preceeding years of maintainability in figure 5.5, we observe that the vast majority of the vehicles rated cost less than $13,000. Vehicles on which no maintenance records were available are denoted by a 0. Significant of the rated vehicles is that they are very evenly distributed across all levels of maintenance scores. This suggests that price alone does not insure the maintainability of an automobile.

3. Characteristics of Size

As an overall expression of size, weight and length show a fairly tight linear relationship as depicted in figure 5.6 . Comparing these two variables across their respective rows indicates that they both appear to manifest similiar relationships with the other parameters. With respect to the size variables of headroom, rear seating, and trunk space, tighter relationships are seen with length. Weight on the other hand has a tighter relationship with the engine displacement parameter (see figure 5.6).

Rear seating and trunk size are both generally increasing relative to length. This observation is what we might expect since a longer external size could reasonably result in larger internal size features. Unusual however, is the factor of headspace, which deviates from the general trend of the other internal size dimensions. As vehicle length increases in figure 5.6, there is a propensity for headspace to encompass a widening range of values. The broadest range is at 106 inches where headspace varies from 1 to 4 inches. As vehicles become even larger, the headspace sizes shifts to the larger values while in general being limited in range from 3 to 5 inches.

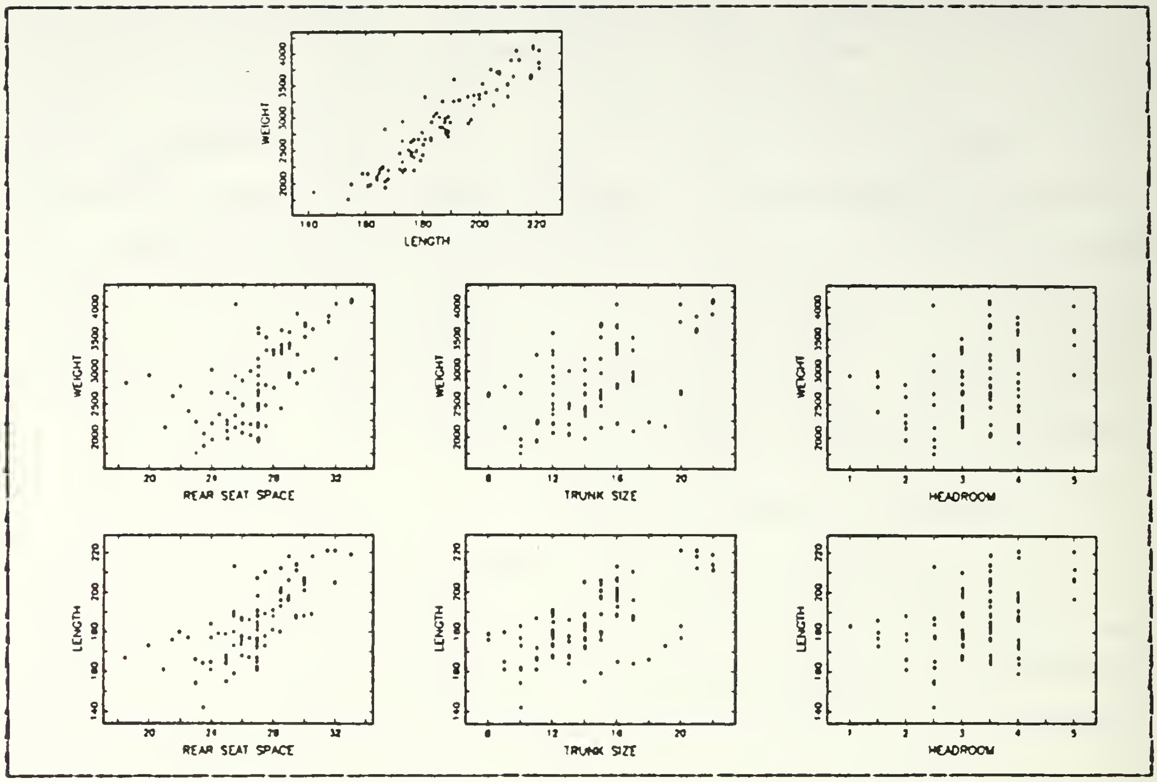


Figure 5.6 Size and Internal Dimensions.

The tendency for increased car weight to be associated with a related increase in engine displacement is another observation we would expect, and is seen in figure 5.7 . Of significance, is that displacement versus weight fall into two distinct types of groups. Vehicles weighing up to 3200 pounds have a very tight increasing linear relationship with engine displacements up to 175 cubic inches. The vehicles of larger weight capacity are seen to be associated with larger engine displacements albeit with a more dispersed cluster of points.

Engine displacement in turn can be seen to have a definite correspondence to the overall automobile categories. A close look at the second plot in figure 5.7 reveals that small cars tend to be banded with engine displacement

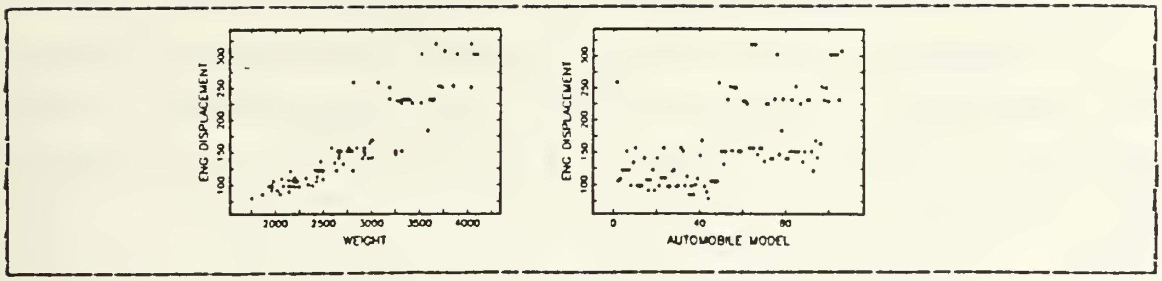


Figure 5.7 Size, Displacement and Vehicle Model.

from 50 to 150 cubic inches. Noticeable is the one small car outlier, identified as the AMC Spirit 6 with a much larger displacement of 258 cubic inches. Medium category cars are fairly evenly distributed in two bands of engine displacement. The lower band spans displacements of 125 to 160 cubic inches while the upper band is tighly spanned from 220 to 260 cubic inches in displacement.

The outliers in the medium car class with significantly larger engine displacements were identified as the Chrysler Cordoba V6, Chrysler Cordoba V8, and Lincoln Continental V8. Almost all of the larger cars are clustered at the 300 cubic inch displacement level with two exceptions. The Buick Electra V6 and Buick LeSabre V6 with displacements of 252 and 231 in3 respectively have lower displacements. Notwithstanding the outliers, vehicle class and engine displacement are very correlated. Overall, the deviations of the outliers in figure 5.7 have an interresting property. They are all of American manufacture and either deviate up or down one engine displacement group. These traits suggest that these vehicles may have previously been in a different size class and changes in some other characteristic features resulted in their being moved up or down an automobile class.

4. Vehicle Performance

Consumers should have a particular interest in the fuel efficiency characteristics of automobiles. Not surprising is the trend shown in figure 5.8 that correlates fuel efficiency in the city with fuel efficiency on the highway. A comparison of the remaining variable plots for these efficiency parameters shows identical relationships in all cases. The original data structure could probably exclude one of these fuel efficiency variables without loss of information if we needed to condense the data.

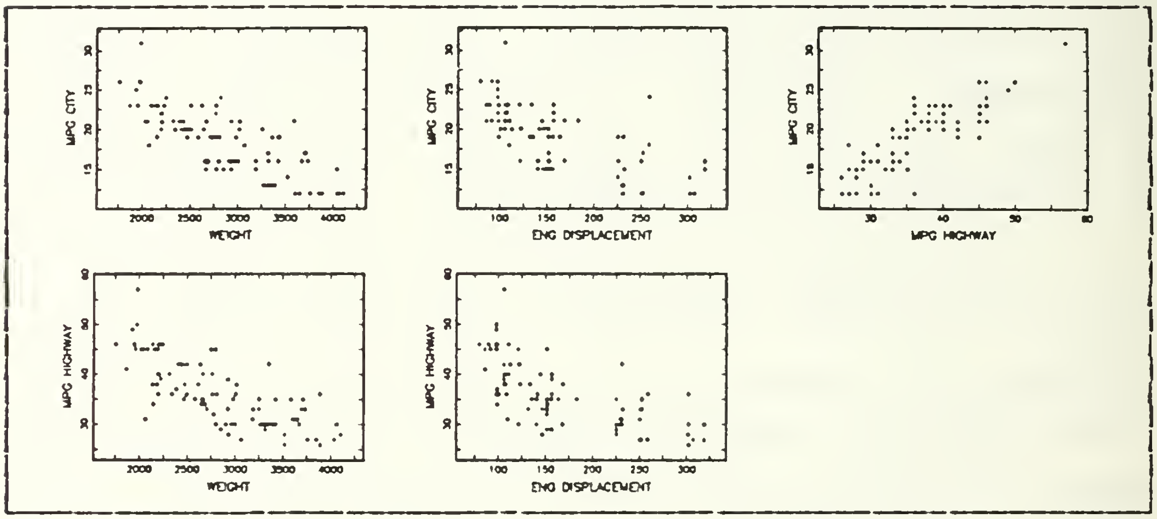


Figure 5.8 Fuel Efficiency, Weight, and Displacement.

The inverse relationship of fuel efficiency to vehicle weight should not be unexpected and confirms our intuition in this regard (figure 5.8). High fuel efficiency, low weight, and smaller engine displacements are all associated.

As previously mentioned, price alone is not an indicator of automobile maintainability. An interesting observation however can be drawn from the relationship exhibited

between repair record from one year to the next. The plot of Repair 81 versus Repair 82 in figure 5.9 indicates a strong positive correlation between the two. In almost all instances the maintainability does not change for better or worse by more than one level. Furthermore, the number of automobiles that improved, deteriorated, or did not change in terms of maintainability are approximately equal.

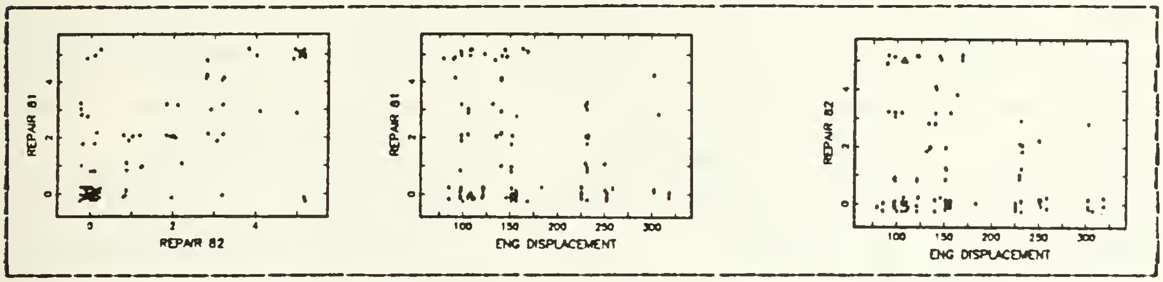


Figure 5.9 Maintainability of Automobiles.

Repair when compared to engine displacement reveals that the predominant number of rated vehicles (i.e vehicles for which repair data was available) contained the smaller displacement engines. This concentration of better maintenance values at the lower displacement level suggests that smaller engines have better maintenance records.

D. ANALYSIS WITH ENHANCED DISPLAY

1. General

The analysis of the basic draftsman display revealed a wealth of features pertaining to the individual variables within the data. One distinct feature evident is the potential relationship between foreign and American manufactured automobiles. While not an original parameter of the data, the plots of price versus weight and price versus model indicates that this influence warrants closer scrutiny.

In the basic display the overlapping of maintenance values was alleviated with jittering. The array of plots dealing with headroom space suggests that this variable also should be treated likewise. The remaining automobile characteristics do not suffer from any significant problems with overlapping values.

Based upon the preliminary analysis, an enhanced display was generated for subsequent evaluation. The redundancy of the two fuel efficiency variables was resolved by eliminating miles per gallon on the highway. The remaining variables were reordered to place those with similiar relationships in closer proximity. The enhanced display also introduces a new discrete category variable, location of manufacture. A value of 1 for this variable corresponds to those automobiles produced in America, those vehicles produced overseas under an American brand name are denoted with a 2, while foreign models are assigned a 3.

The introduction of location of manufacture dramatically portrays some very evident dichotomies which exist between foreign and American made automobiles. In general, the array of plots consisting of these parameters indicates a very different orientation on the part of the respective manufactures in their approach to the automobile market.

The potential for transforming the data through transformations was considered. Transforming engine displacement with a log transform slightly straightens the plots containing this variable with respect to some of the other size parameters as seen in figure 5.10 . This reexpression however does not really enhance the description of the data and hence was not included in the final display.

The complete enhanced draftsman display may be found in Appendix B. Isolated portions of this display will be reproduced within this section of the text for clarity.

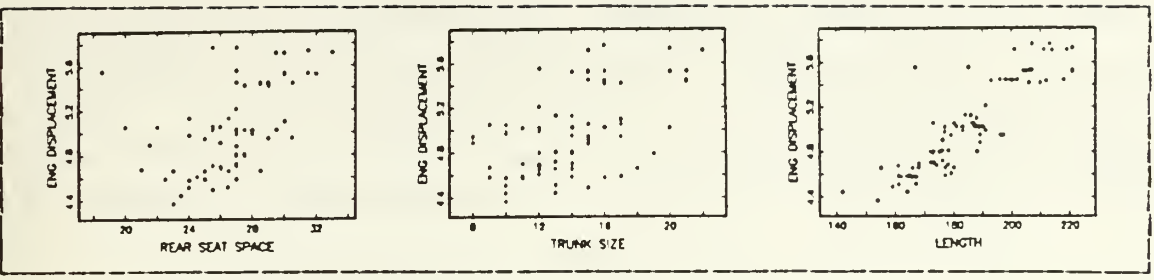


Figure 5.10 Log Transformation of Engine Displacement.

2. Price

The majority of American and foreign cars are primarily priced below $12,000. Extremely visible is the large grouping of American models in the neighborhood of $9000 to $11000 as depicted in figure 5.11 . American model prices beyond this level tend to increase in a rather uniform fashion of $2000 increments, up to $24000. In contrast, foreign car prices are fairly uniformly distributed in the region of $5000 to $15000 with subsequent price hikes in larger increments of $5000, to the maximum level of $35000.

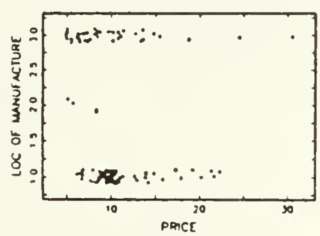


Figure 5.11 Location of Manufacture and Price.

3. Size

In general, automobiles of American and foreign origin fall within two distinct size ranges (figure 5.12).

In terms of the major dimensions of length and weight, the plot arrays provide some distinguishing features contrasting location of manufacture. Not very surprising is that American vehicles tend to the longer and heavier side while foreign manufactured cars tend to be shorter and lighter.

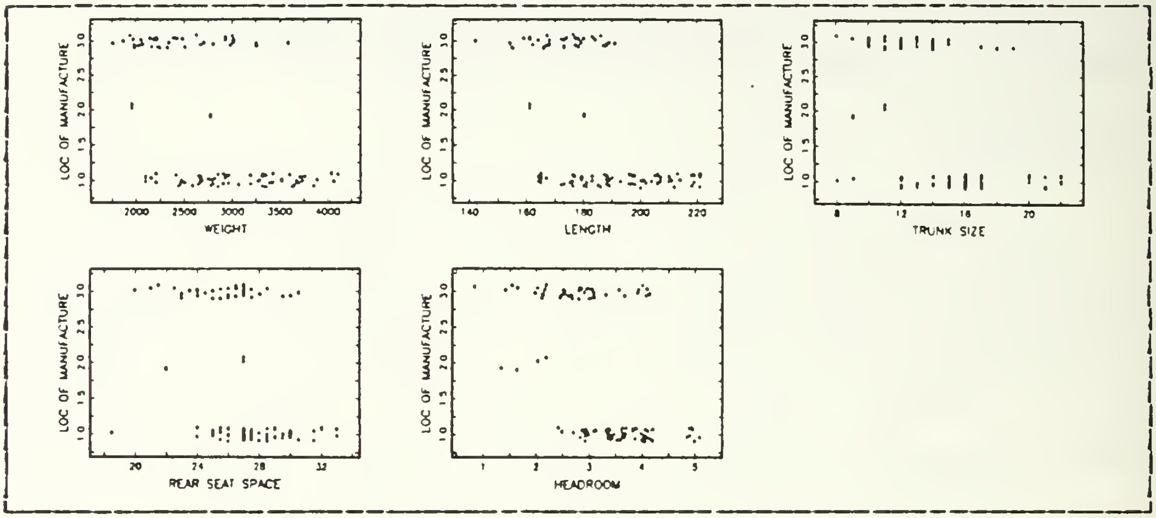


Figure 5.12 Location of Manufacture and Size.

In view of the propensity for foreign cars to be shorter than the American counterparts, the plots of the related inner size characteristics shown in figure 5.12 is somewhat unexpected. An evaluation of rear seating space, trunk size, and headroom shows that the distribution of values for foreign produced vehicles is slightly shifted to the smaller dimensions in contrast to the respective American distributed values. This is consistent with observations noted in the basic display. What is unexpected, is that the differences based upon the shifts is much smaller than we might expect given the prevalent difference in length distributions between American and foreign cars.

In conclusion, the rear seat spaciousness of American cars is fairly evenly distributed between 25 and 30

inches with the large models being outliers at 32 inches. The foreign models are more widely dispersed from 20 to 29 inches. Thus at the upper end of the spaciousness scale there is actually only a three inch advantage by the largest of the American models.

The characteristic of headroom denotes a similiar relationship as that observed for rear seating space (figure 5.12). Again, approximately 50% of the foreign car headroom values fall within the same distribution range (2 1/2 to 4 1/4 inches) as that of the vast majority of the American models.

The differences between trunk sizes seen in figure 5.12 is a bit more acute in that the foreign models range from 10 to 14 cubic feet, while the American cars are skewed toward 12 to 14 cubic feet. Clearly, in spite of the distinct length differences between foreign and American produced cars, the differences in internal dimensions is much subtler and smaller than we might have originally suspected. The foreign cars, although smaller in overall length , have approximately the same internal size features as all but the very largest American made cars. It is also interesting to note that the American sponsored but produced overseas models tend to exhibit the characteristics of the foreign models.

4. <u>Performance</u>

The distribution of the fuel efficiency characteristics of American and foreign automobiles appears to be the inverse of their weight (figure 5.13). The heavier American cars tend to be evenly distributed between 9 and 20 mpg with only three outliers extending beyond 25 mpg. The lighter foreign cars, while ranging from 15 to 28 mpg, are rather tightly grouped between 20 and 25 mpg. The outlier in this case is at the extreme range of 33 mpg. In terms of

fuel efficiency, the foreign vehicles certainly dominate this attribute of performance.

Perhaps the most revealing plots in the enhanced display are those of the repair records. In both recorded years the American models are rather evenly distributed from poorer than average (1) to average (3) maintainability ratings. A better than average rating (4) was achieved only four times over both years. In extreme contrast, the foreign models during both years show a tendency toward the much better than average maintainability rating (5). As in the characteristic of fuel efficiency, the foreign models dominate this performance variable of maintainability.

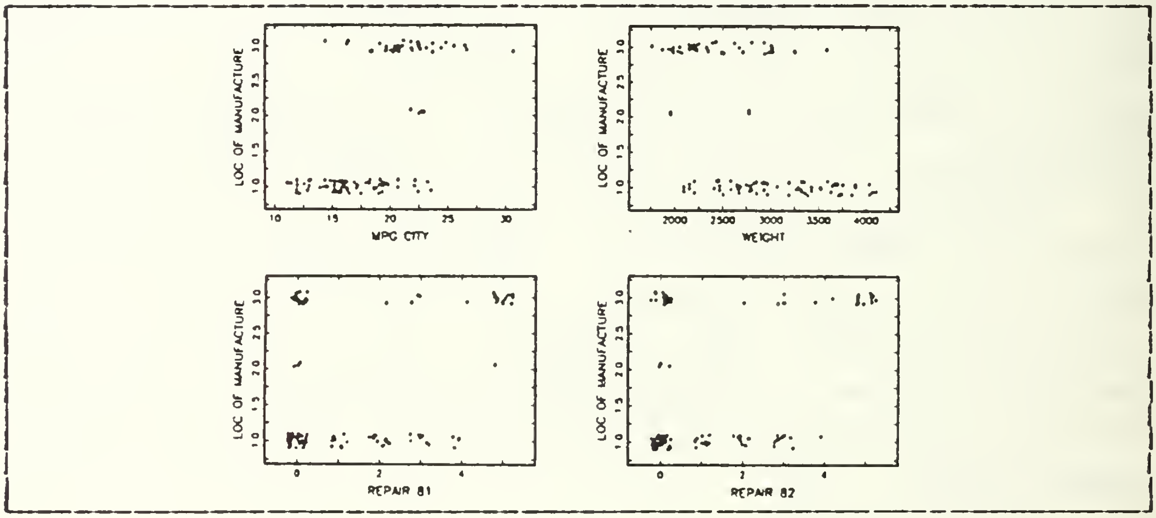


Figure 5.13 Location of Manufacture and Performance.

E. CONCLUSIONS

The car data is an excellent example of how the Draftsman display can be used to describe a data set. The various parameters associated with automobiles can be very confusing to the consumer. No one single parameter can be

selected as an overall measure of what constitutes a "best" automobile. What one consumer may find desirable, another consumer may find unacceptable. Thus, describing or modeling the data with more formal statistical techniques such as linear regression is not very applicable. The Draftsman display enables the user to observe the multivariate affects of each of the various parameters. Based upon the selection of one or more parameters, the user can determine the impact relative to other parameters.

VI. AN ANALYSIS OF CONTRACT DATA

A. INTRODUCTION

A graphical analysis is presented using the Draftsman display on data collected concerning selected Naval contracts signed during the period 1949 through 1963. This chapter explores the general descriptive qualities of eleven categories of contractual information relative to the performance of the contracts.

The data originally was analysed in a Thesis completed in 1973 [Ref. 7], through regression and analysis of variance techniques. Significant in this study was the conclusion that there was no clear method of describing the relationships between contract parameters and the subsequent performance of the contracts. It is this authors opinion that the analysis failed because the use of linear regression alone is not sufficient to adequately describe the relationships present in the data. The analysis presented based upon a Draftsman display suggests that this method of exploratory data analysis reveals a variety of relationships do exist describing contract performance relative to the contractual parameters.

B. THE CONTRACT DATA

The data consist of 177 contracts which comprise all Naval aircraft and missile fixed-price incentive contracts completed during the period 1949 through 1963. The data as provided by the Naval Material Command encompasses 11 parameters as follows:

1. Deviation from target cost (percent).
2. Months to complete contract (months).

3. Target profit of manufacturer (percent).
4. Sharing ratio (percent).
5. Ceiling price (percent of target price).
6. Target cost of contract (millions of dollars).
7. Number of items produced in the contract.
8. Number of contracts let that year.
9. Year the contract was signed (see table III).
10. Contractor awarded contract (see table III).
11. Type of system (see table III).

TABLE III

Description of Variable Coding

Codes for variable 9
YEAR SIGNED

1 = 1949
2 = 1950
⋮
15= 1963

Codes for variable 11
SYSTEM TYPE

1	Utility Airplane
2	Combat Airplane
3	Missile
4	Blimp
5	Helicopter
6	Drone
7	Airborne Equipment

Codes for variable 10
MANUFACTURER

1	Beech	7	Hiller	13	Ryan	19	Philco
2	LTV	8	Kaman	14	Sikorsky	20	Maxson
3	Corvair	9	Martin	15	Bell	21	Northrop
4	Douglas	10	McDonald	16	Lockeed	22	Raytheon
5	Boeing	11	N. American	17	Bendix	23	Aerojet.
6	Grumman	12	Vertol	18	Gen Elect.		

C. THEORY OF FIXED-PRICE INCENTIVE CONTRACTS

The concept behind fixed-price incentive (FPI) contracts is that they are intended to be used in the development,

management support, and production of items in which the uncertainty of cost is too great to allow a firm-fixed price (FFP) contract attractive to bidders. In theory, the FPI contract should motivate control of costs by rewarding the manufacturer with a greater profit level as costs are reduced below the negotiated target cost.

The incentive feature of the FPI contract should influence the contractors to effectively manage cost associated decisions in a manner beneficial to profit. This in turn should result in a favorable cost outcome to the government as well. This mutually favorable outcome is communicated in the form of the sharing ratio which establishes the amount of money which will be returned to the contractor for every dollar saved below the target cost of the program. For example, a 75/25 sharing ratio returns 25% of every dollar saved to the contractor while reducing the governments expected cost by .75 dollars. The higher the percent returned, the greater the potential for gain or lose to the contractor. Hence, lower sharing ratios reflect a greater degree of financial risk to the contractor.

The ceiling price of a contract is a control measure to avoid excessive cost overruns to the government. The ceiling price establishes the maximum amount of cost which will be paid by the government. When final cost exceeds the ceiling cost, the difference must be borne out of pocket by the contractor as a loss. Cost outcomes which fall between the negotiated target and ceiling values result in a break even venture to the manufacturer.

D. PRELIMINARY ANALYSIS

1. General

A Draftsman display of the eleven contractual parameters was generated for preliminary analysis. Most

noticeable was that nine of the eleven variables consisted of discrete values which resulted in substantial overlapping of plotted points throughout the display. The plots containing the parameter of number of items produced indicates a problem in scaling. Contracts range in size from 40 to 1400 items. This problem as shown in figure 1.1 , is caused by the eight extreme outlier contracts containing in excess of 800 items. The compression of the remaining majority of the contracts into a very small segment of the plots would prevent observations of any meaningful value.

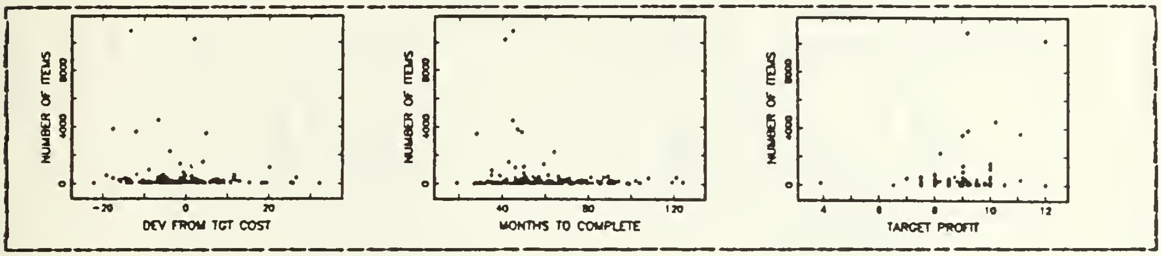


Figure 1.1 Number of Items per Contract.

A subsequent Draftsman display was generated to alleviate the problem of overlap as well as the scaling of number of items per contract. To take care of the overlap problem all variables except deviation from target cost, target cost, number of items, and manufacturer were jittered. To take care of the scaling problem, a log transformation was used on the variable of number of items.

The Draftsman segments generated with enhancements were reduced and are shown in figures 1.2 through 1.5 . For convenience and clarity of discussion appropriate plots will be reproduced within the body of the chapter text. The original display segments may be seen in Appendix C.

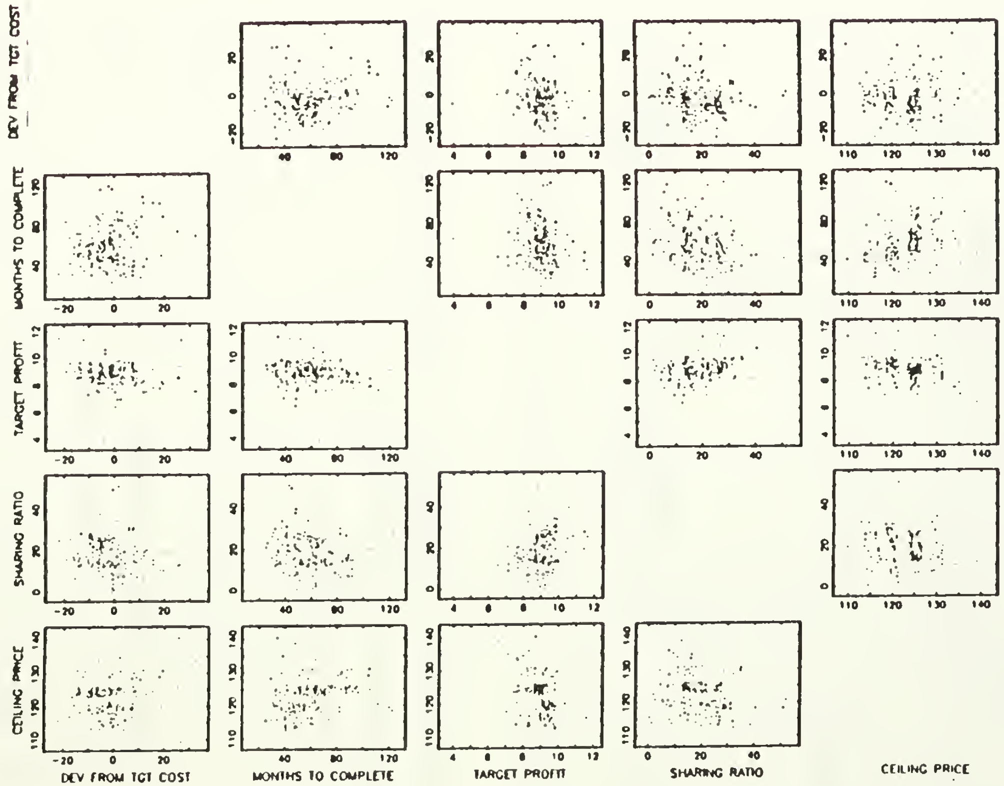


Figure 6.2 Draftsman Segment 1, Contract Data.

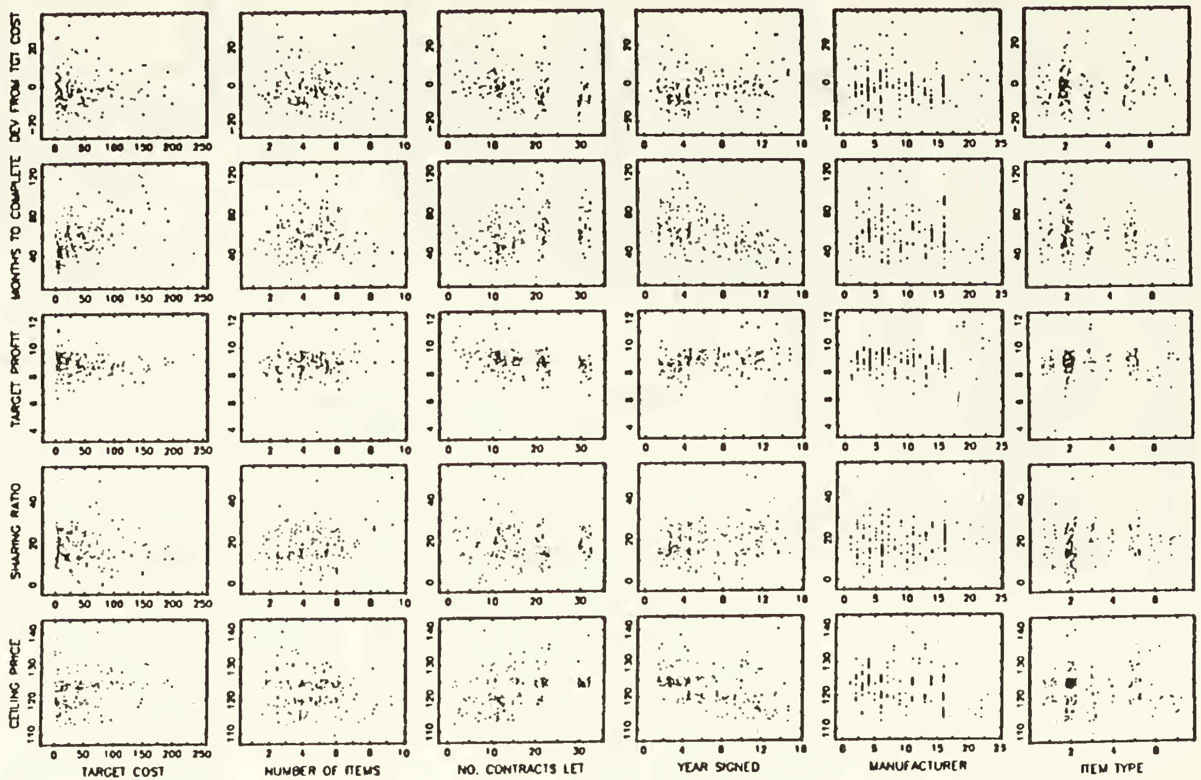


Figure 6.3 Draftsman Segment 2, Contract Data.

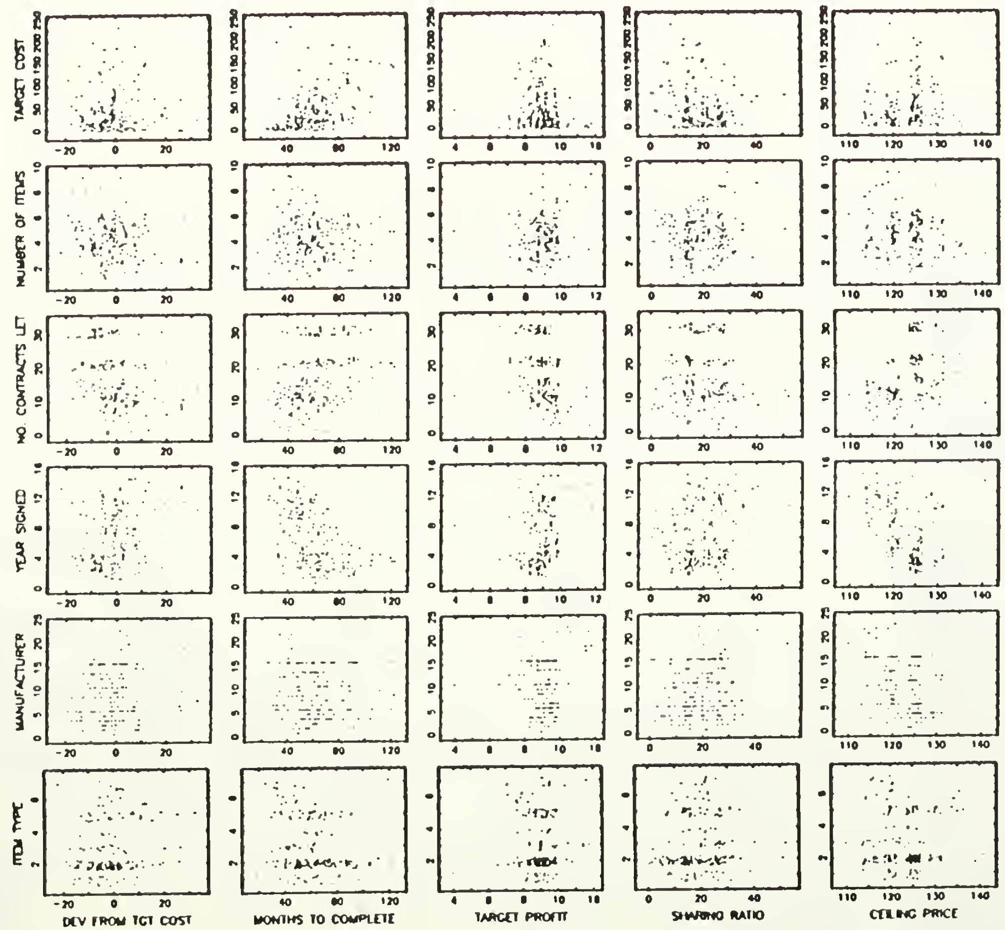


Figure 6.4 Draftsman Segment 3, Contract Data.

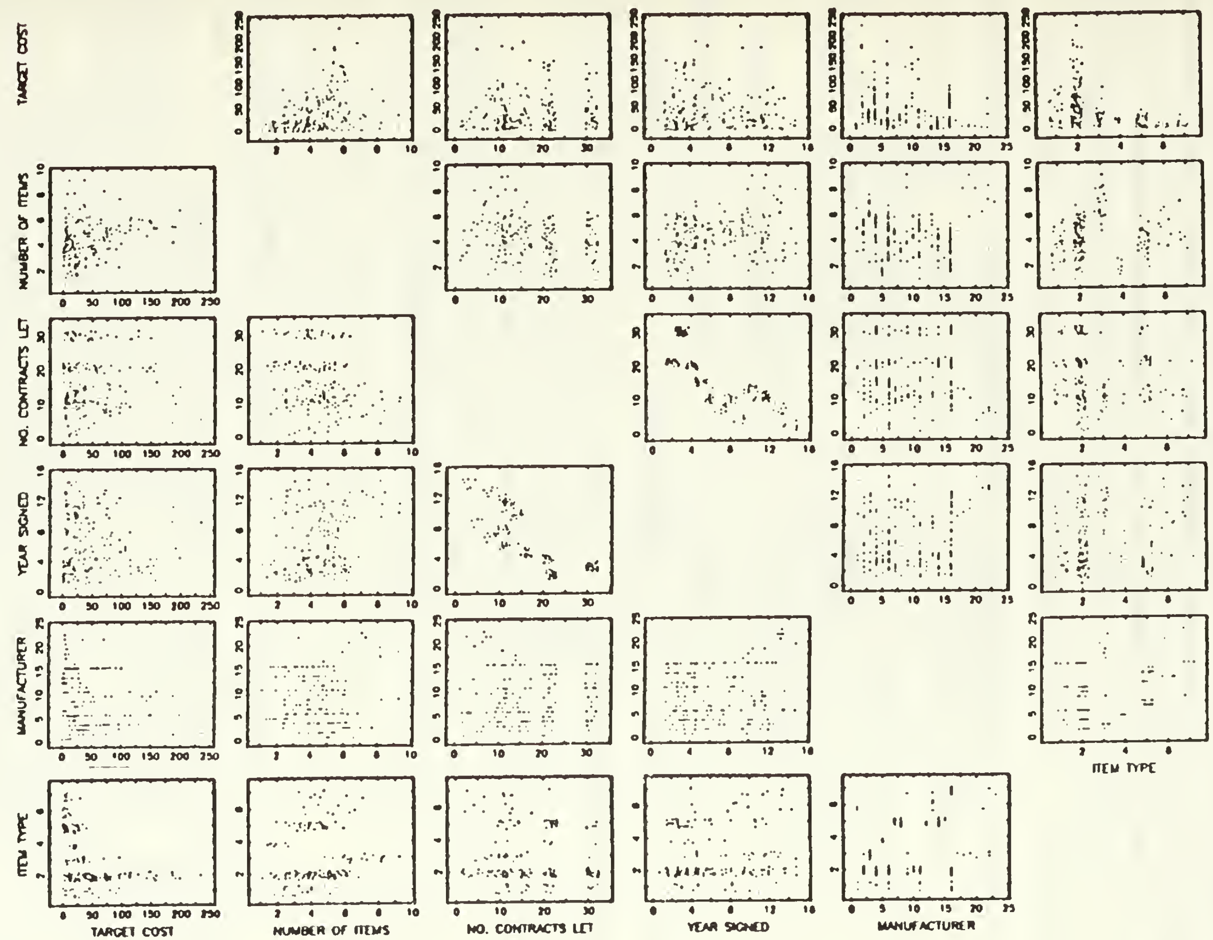


Figure 6.5 Draftsman Segment 4, Contract Data.

2. Characteristics of Size

a. Volume of Contracts

As a preliminary expression of size, the number of contracts per year provides an estimate for the volume of contracts let during a particular period of time. We might suspect that a low volume of contracts would create a more competitive atmosphere among manufacturers as they attempt to maintain their facilities in a production mode. A high volume of contracts let offers a greater opportunity for manufacturers to select contracts in which they have the greatest amount of expertise and experience. The latter case has a greater potential for controlling costs as well as manufacturer profits. Thus, we might expect that as the number of contracts let increasess, the positive deviations from target cost should decrease.

The plot of cost deviation versus number of contracts seen in the first plot of figure 6.6 does appear to generally support this hypothesis. As the volume of contracts increases there is a tendency for cost deviation to be negative. In fact, the greater the volume, the greater in magnitude the negative cost deviation.

The cost deviations versus volume relationship, when compared over time, also suggest that the time at which the contracts were signed may have additional bearing (see figure 6.6). The rapid increase in contract volume from 1949 to 1951 is characterized by large absolute deviation from target cost (though generally negative). As volume declined from 1951 through 1955, the absolute deviations from target cost can be seen to be much smaller and roughly equally distributed between positive and negative. The subsequent volume increase experienced from 1955 to 1958 also shows a increase in the absolute deviations from target cost (with a fair tendency toward negative deviations). The

me decline from 1958 to 1964 is somewhat more difficult nterpret due to the relatively low volume of contracts. lute deviations do appear to become slightly smaller as me decreases. However, when contrasted to previous s of similiar volume, the deviations while becoming ler, appear to be doing so to a lesser degree than iously. The last three years of the data period, while acterized both by low volume as well as a small absolute ation from target costs, clearly shows a tendency rds positive cost deviation.

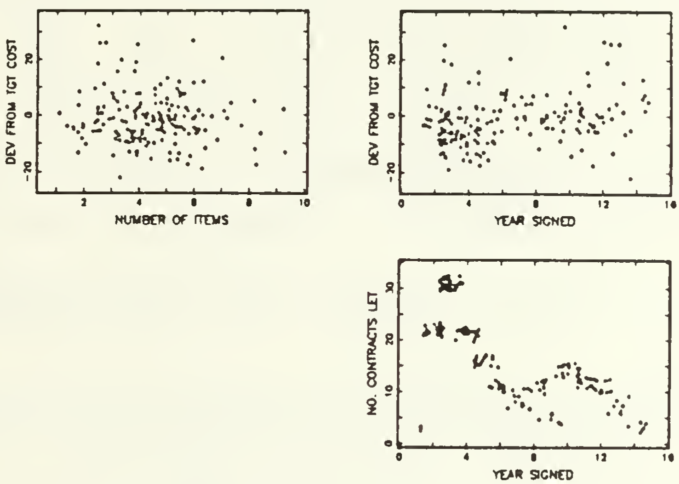


Figure 6.6 Contract Volume.

b. Contract Duration

Based upon production management techniques, we t expect that the duration of a contract would have a tionship to contract performance. Short term contracts e little time for management to adjust production activ- s to maximize the efficiency of operations. As contract tion increases, a greater opportunity is afforded to

contractors to learn by early production errors and make the cost related decisions necessary for control. When contracts duration extends far into the future, difficulties can arise by external economic influences which could not be accurately forecast at the onset.

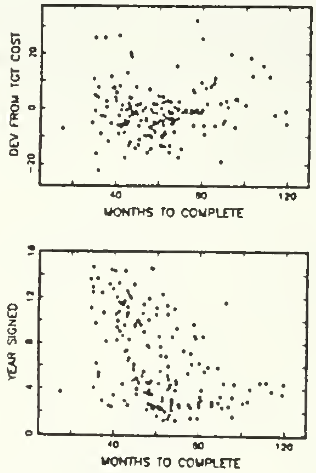


Figure 6.7 Contract Duration and Performance.

The isolated plot of deviations from cost relative to contract duration reproduced in figure 6.7 reveals some interesting features. The contracts of less than 40 months duration exhibit widely dispersed deviations from target cost. For the contracts which lasted between 40 and 70 months the cost deviations exhibit a clear trend toward the negative side. Contracts which exceed 70 months in duration show an increased deviation that is roughly equally split between positive and negative .

The cost deviation characteristics relative to duration noted above appear to hold irrespective of the year in which the contracts were signed (see figure 6.7). Contracts of less than 40 months duration as well as those between 40 and 70 months are fairly equally distributed

across all years of the data period. Whether contracts which exceed 70 months in duration are effected by the year signed is not determinable since all but one of these contracts occured during the first five years.

c. Target Cost of Contracts

The target cost parameter provides a measure of the financial size of the contracts let. The plot of this variable with respect to deviation from target cost is shown in figure 6.8 . The greatest absolute deviation from target cost can be observed when target cost is less than 100 million dollars. In this region, the deviations tend to be negative but only by a slight numerical margin. The contracts which exceeded a target cost of 100 million dollars are clearly seen to exhibit a smaller absolute deviation from the target cost. These contracts are further characterized by generally favoring a negative cost deviation.

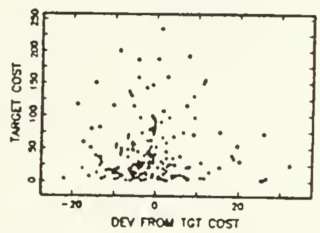


Figure 6.8 Target Cost and Performance.

3. Incentive Measures

a. Sharing Ratio

The sharing ratio establishes the amount of money which will be returned to the contractor for every

dollar saved in cost. This return reflects the potential for profit to the manufacturer. The higher the sharing ratio, the higher the potential gain. This relationship appears to be echoed in the plot of these two variables as seen in figure 6.9 . As the sharing ratio increases so does the relative expected profit level of the manufacturer. The risk factor associated with contract duration can also be observed in the sharing ratio. As contract duration increases the potential for influence by other external economic parameters can less accurately be forecasted. The general decline of sharing ratios as contract duration increases can be seen in figure 6.9 . This decline may be a sign of the contractor's willingness to accept a lower maginal profit position in order to decrease risk.

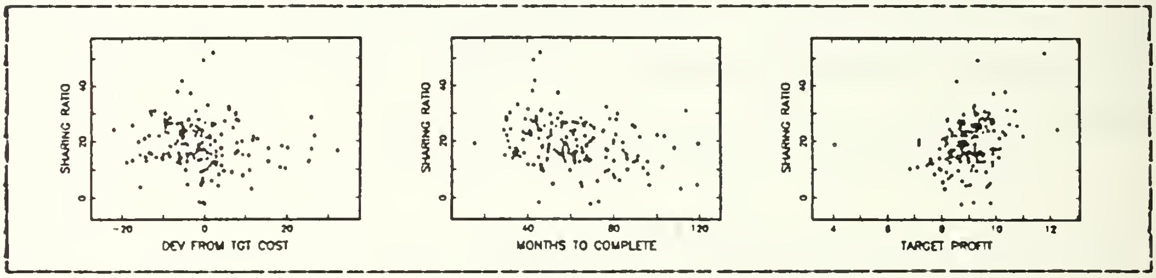


Figure 6.9 The Incentive of Sharing Ratios.

The most striking observation about the sharing ratio is that when evaluated with performance, little if any relationship can be determined. We can see no clear indication that any particular ratio can be associated with a favorable (negative) cost deviation. This lack of relationship is significant in that the sharing ratio is supposedly a major incentive feature of fixed-price incentive contracts. A determination that sharing ratios are an insignificant parameter suggests that this method of contracting might warrant further analysis by the

government. It may be interesting to note that a Rand study of over 400 Air Force contracts resulted in a similiar finding that the sharing ratio was insignificant with respect to the final outcome of contracts [Ref. 8 p.38].

b. Target Profit

The plots of negotiated target profit level in figure 6.10 reveal similiar characteristics when compared to the contract size parameters. Very evident is that target profit in general tends to revolve around the 9% level. Given the time period during which these contracts were performed, 9% represents a rather lucrative profit level.

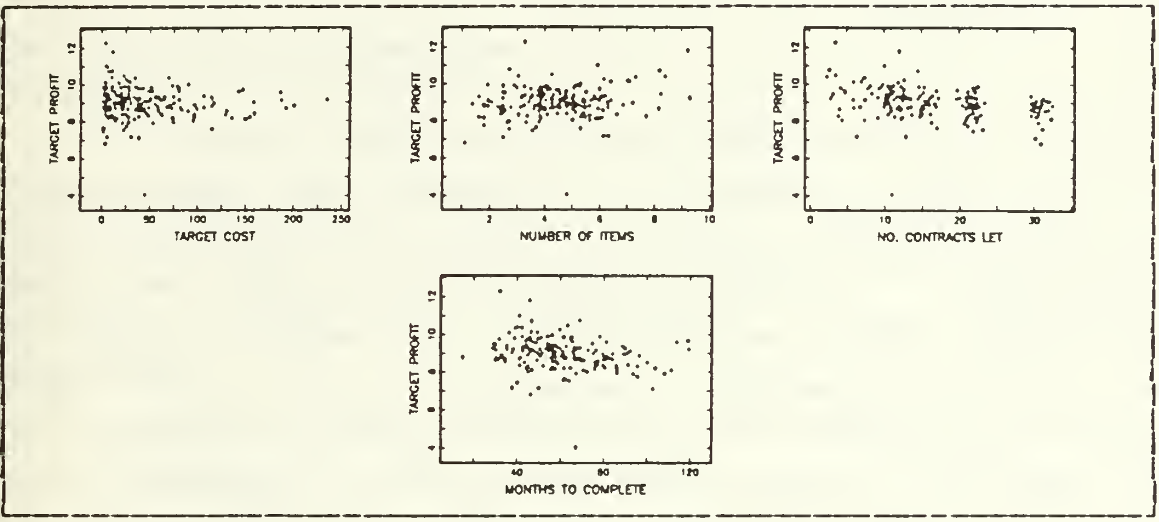


Figure 6.10 Target Profit and Size.

Profit versus target cost is observed to be funnel shaped with the greatest deviations in profit at the lower end of target cost. As target cost increases the variance of profit becomes smaller while stabilizing about the 9% level (figure 6.10). A similiar though looser relationship exists with the number of items produced. Target profit deviates imperceptibly more initially and generally

tends to fluctuate about 2 percentage points above and below the 9% level. The 9% profit level remains firm regardless of the number of items.

Target profit plotted with duration and volume shows a slight decline in expected profit levels as either duration or volume increase as depicted in figure 6.10 . In the former situation this may be a willingness to trade-off a lower profit level for the security of a longer term production operation. The inverse relationship between profit and volume suggests that during low volume periods contractors are attempting to make the most out of the few contracts available. Conversely when contract volume is high the expected profit level declines to about 9%. The conclusion might be drawn that with more contracts available, the contractors are willing to accept a slightly lower profit level per contract since the opportunity is greater to win multiple contracts during high volume periods.

As a measure of performance the target profit may lack significance in determining the deviation from target cost. The plot of target profit versus deviation from target cost shown in figure 6.11 does not suggest a describable relationship. The majority of the manufacturers tended to negotiate about a 9% profit level. An anlysis of the eight most deviant outliers from this characteristic reveals that seven of these outliers were by manufacturers with these contracts being their sole participation during the entire 15 year period. The comparison of target profit, deviation from target cost and system type is also significant with respect to the outliers. While there is nothing notable about their target cost, the eleven most unfavorable cost outcomes (positive cost deviations) correspond solely to three system types. These are combat aircrafts, missiles, and helicopters denoted by item types 2, 3, and 5 respectively in figure 6.11 .

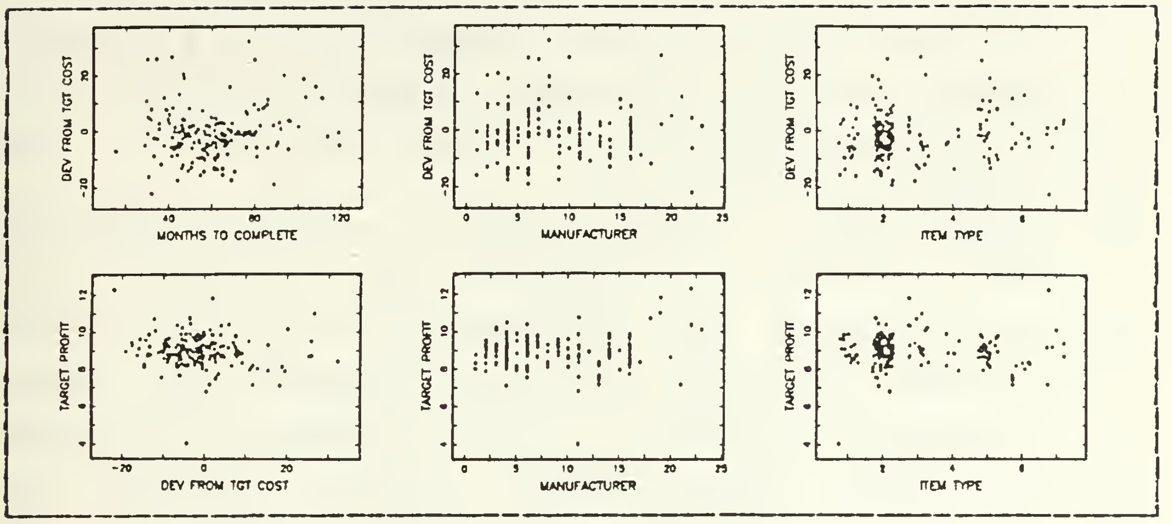


Figure 6.11 Target Profit and Performance.

E. PRELIMINARY CONCLUSIONS

The analysis and discussion presented reveals that an abundant amount of information is visible in the Draftsman display of the contract data. Further, there are indications that there are relationships between the contractual parameters and contractor performance. Briefly summarized these are:

1. As the volume of contracts let increases there is a tendency for the contracts to result in a negative cost deviation (favorable to the government).
2. Over the 15 year period as volume changed from year to year there appears to be a related reaction relative to contractor performance. Periods denoted by an increasing volume are reflected with an increase in cost deviations. When volume declines a related decline in cost deviations can also be observed but at a more cautious rate.
3. Contract duration as related to cost deviation might better be described in terms of short, medium, and

long term contracts. The most favorable contract duration appears to be between 40 and 70 months. This relationship is fairly consistent regardless of the year in which the contracts were signed or the volume of contracts let.

4. Fluctuations in cost deviations tends to stabilize when the contracts contain more than 50 items.
5. As the target cost increases there is a greater tendency for negative cost deviation to occur. Contracts in excess of 100 million dollars in particular resulted in predominantly favorable outcomes.
6. The sharing ratio as a traditional incentive measure of a FPI contract may lack merit. No relationship can be observed between this parameter and performance.
7. No obvious relationship can be noted between target profit levels and contract performance.
8. The ultra high technology systems of combat aircraft, helicopters and missiles exhibit the greatest potential for adverse performance.

F. ADDITIONAL CONFIRMATORY ANALYSIS

1. General

The preliminary analysis using a single iteration of the Draftman display revealed a variety of interesting relationships between contractual parameters. Certainly other relationships exist which have not been discussed. As an exploratory data analysis tool the Draftsman display enables the user to look at the data at almost any level of detail desired. Subsequent displays can be generated on various subpopulations such as each of the manufactures to gain greater insight to their performance behavior. It is this versatility in exploring data sets which enables the user to

rapidly process large amounts of data in order to gain a feeling for the interactions involved.

The use of the Draftsman display can also assist the user in the application of more formal statistical approaches. The use of such techniques as regression analysis provides a confirmatory measure to the exploratory indications viewed in the display.

The blind application of some statistical packages without first looking at the data can result in erroneous or misleading conclusions. This is particularly true of large data sets where misplaced decimal points, formating errors and other related problems may not be easy to detect. The visual nature of the Draftsman display can assist in identifying these problems as well as aid in selecting appropriate variable selections on which to initiate formal analysis.

2. Cost Deviation Over Time

One major question which cannot be readily answered with the Draftman display of contract data is the relationship of cost deviations over the entire contract period. The scatter plot of percent cost deviation versus year signed reveals a wide dispersal in cost deviations with no clear visual trend apparent (figure 6.12).

A least squares linear model was selected in order to determine if for all manufacturers a trend exists relating cost deviations to the year in which contracts were signed. The results of this indicates that in fact an upward trend in cost deviations did occur from 1949 through 1964 (figure 6.12). The computed t-value of 3.3 is quite significant and indicates that the probability that the value of the coeficient B(1) was actually zero is much less than .05. The slope of the regression line is .580 with the lower and upper confidence intervals .232 and .927 respectively. This also clearly supports the upward trend of cost deviations.

TERM	COEFFICIENT B	SIGMA(B)	B/SIGMA(B) T	CONFIDENCE INTERVAL LOWER	UPPER
B0	-4.920	1.318	-3.740	-7.517	-2.323
B1	.580	.178	3.301	.233	.927

THE THEORETICAL VALUE FOR T AT THE 0.05 LEVEL AND 175 DF = 1.974

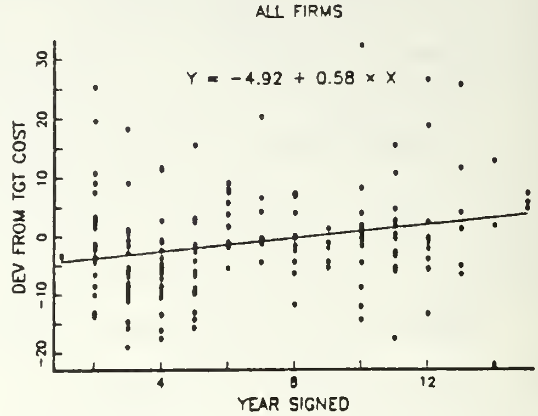


Figure 6.12 Contract Performance Over Time.

While an overall increasing trend in cost deviations is evident, the performance of the individual manufactures involved might reasonably be expected to differ. The generation of Draftsman displays for each manufacturer would provide the starting point for comparison. While the entire displays are not presented, the scatter plots of two major contractors (Grumman and Lockeed), indicate how different performance results relate in the general cost deviation picture.

In applying the linear regression model to Grumman as seen in figure 6.13, cost deviations rose rapidly over the time period. This rise is much faster than that seen in figure 6.12 for all the firms in general. From a government perspective this might suggest that a closer scrutiny of this company's activities might be warranted.

The application of the regression model to Lockeed as seen in figure 6.14 presents a very different picture. In this instance a cubic fit rather than a straight line is

TERM	COEFFICIENT B	SIGMA(B)	B/SIGMA(B) T	CONFIDENCE INTERVAL LOWER	UPPER
B0	-12.343	3.191	-3.858	-18.962	-5.724
B1	1.573	.413	3.813	.717	2.429

THE THEORETICAL VALUE FOR T AT THE 0.05 LEVEL AND 22 DF = 2.074

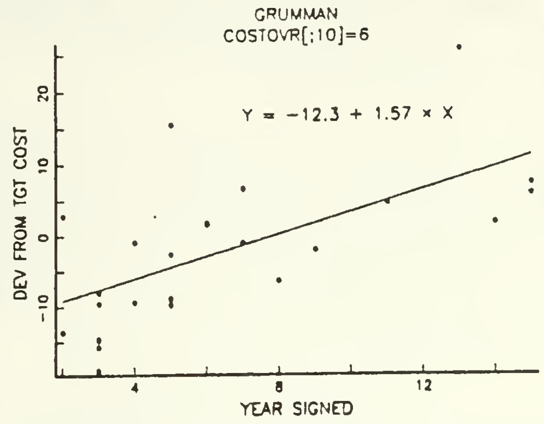


Figure 6.13 Grumman Contract Performance.

more appropriate in describing the relationship of cost deviations over time. Cost deviations appear almost cyclical. Particularly noteworthy is the difference between Grumman and Lockeed during the last four years of the data period. Grumman cost deviations continue to rise while Lockeed experienced a sharp decline in deviations. Quite likely there are external considerations which are influencing cost for each of the manufacturers.

TERM	COEFFICIENT B	SIGMA(B)	B/SIGMA(B) T	CONFIDENCE INTERVAL LOWER	UPPER
B0	13 913	6 802	2.045	¯ 073	27 898
B1	¯10.832	3 629	¯2.985	¯18 293	¯3.371
B2	1.857	.556	3.339	.714	3.001
B3	¯.089	.025	¯3.489	¯.141	¯.037

THE THEORETICAL VALUE FOR T AT THE 0.05 LEVEL AND 26 DF = 2.056

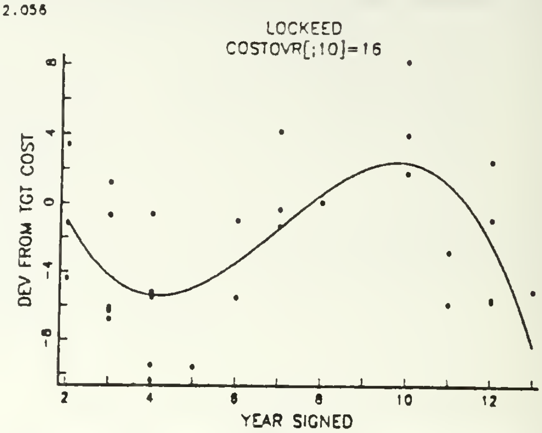


Figure 6.14 Lockeed Contract Performance.

APPENDIX A

DRAFTSMAN COMPUTER CODE

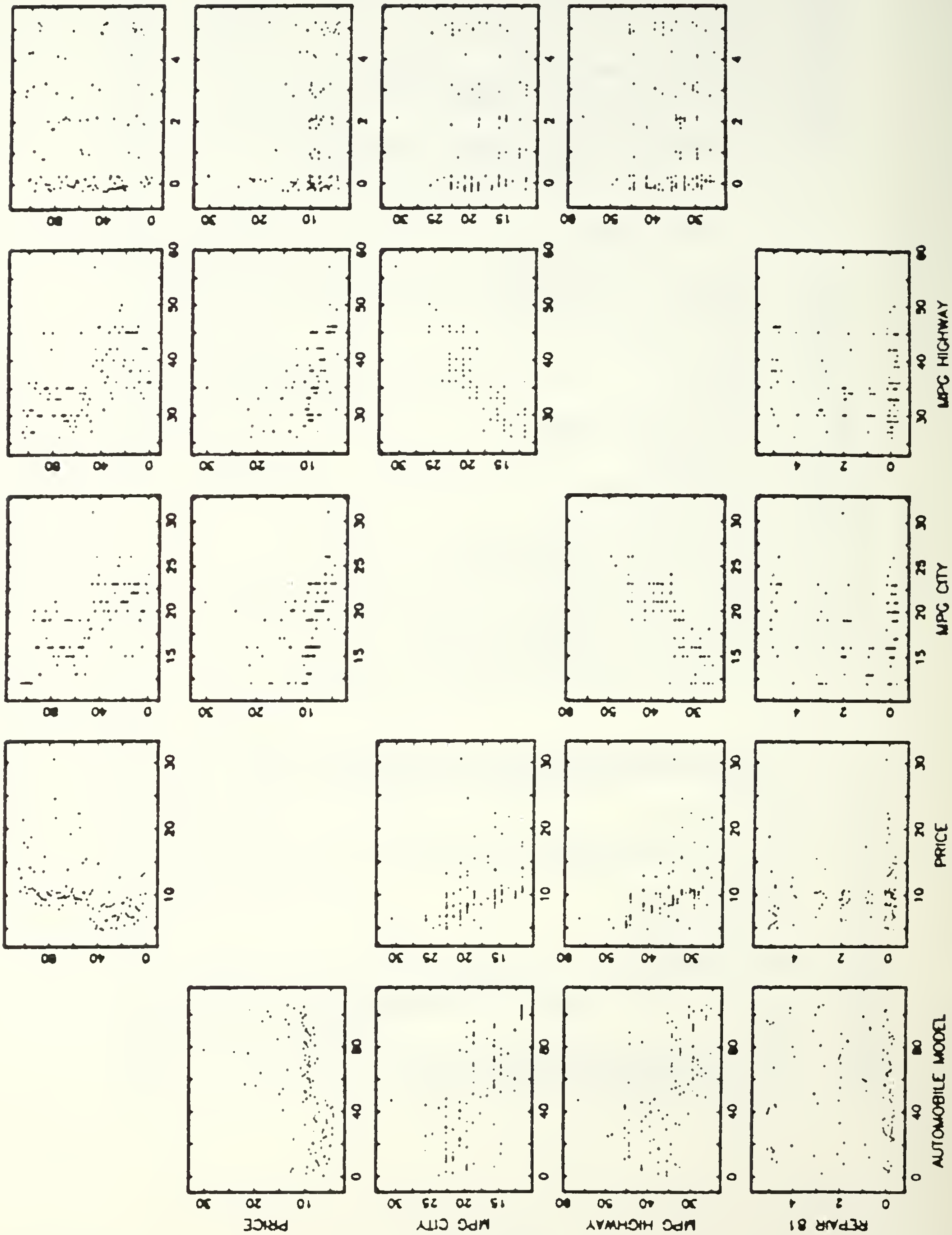
```
  ∇ DRAFTSMAN;DATA;NCOL;TR;TC;PI;R;C;Y;TN;T2N;XAXIS;YAXIS;X;
    TX;LY;TY;N POSN;LX;
  ADMIN
  NCOL←¯1↑(⍴DATA)
  JITTER
  TRANSFORM
  TR←¯5
LOOP4:TR←TR+5
  TC←¯5
LOOP3:TC←TC+5
  PI← 0.1 0.82 0.25 0.97
  R←0
LOOP2:R←R+1
  C←0
  Y←DATA[;(TR+R)]
LOOP1:C←C+1
  X←DATA[;(TC+C)]
  →((TR+R)=(TC+C))/SKIP
  POSN←PI+((0.18 ¯0.18 0.18 ¯0.18)×((C−1),(R−1),(C−1),(R−1))
  XAXIS←N[(TC+C);]
  YAXIS←N[(TR+R);]
  →((C=1)∧((R=5)∨((TR+R)=NCOL)))/GRAPH
  XAXIS←''
  →(C=1)/GRAPH
  XAXIS←N[(TC+C);]
  YAXIS←''
  →((R=5)∨((TR+R)=NCOL))/GRAPH
  XAXIS←YAXIS←''
GRAPH:MINMAX
  RUN BASIC
SKIP:→(((TR+R)≥(NCOL))∧((TC+C)≥NCOL))/END
  →((C<5)∧((TC+C)<NCOL))/LOOP1
  →((R<5)∧((TR+R)<NCOL))/LOOP2
END:PAUSE
  ERASE
  →((TC+C)<NCOL)/LOOP3
  →((TR+R)<NCOL)/LOOP4
  ∇
```

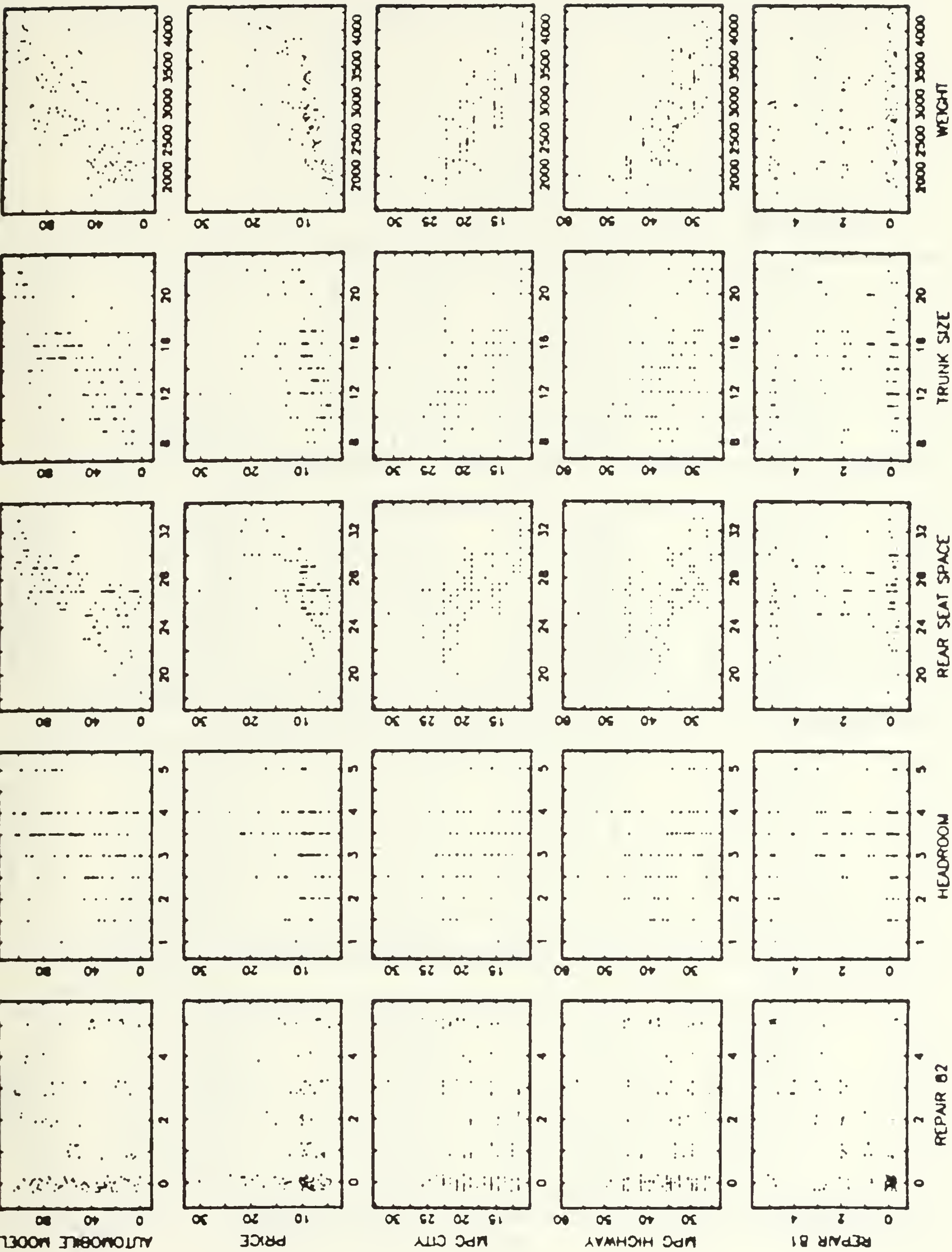
```
∇ ADMIN;RZ1;RZ2;RZ3
  'IS YOUR TWO DIMENSIONAL DATA SET '
  'ALREADY LOADED IN THIS WORKSPACE?'
  '( Y OR N )'
  RZ1←⎕
  →(RZ1='N')/LA
  'WHAT IS THE NAME OF THE DATA SET'
  DATA←⎕
  →LB
 LA:'TO HAVE YOUR DATA READ INTO '
  'THIS WORKSPACE FROM A CMS FILE'
  'ANSWER THE FOLLOWING QUESTIONS'
  DATA←CMSREAD
 LB:'DO YOU DESIRE ALL OF THIS DATA'
  'TO BE PRESENTED IN THE DRAFTSMAN'
  'DISPLAY OR JUST A SUBSAMPLE OF IT?'
  'ENTER (ALL OR SUB)'
  RZ2←⎕
  →(RZ2='ALL')/LC
  SUB DATA
 LC:'DO YOU HAVE A TWO DIMENSIONAL'
  'ARRAY OF NAMES FOR THE DATA'
  'WHICH IS TO BE DISPLAYED? ENTER (Y OR N)'
  RZ3←⎕
  →(RZ3='Y')/LD
  LABELS DATA
  N←NAMES
  →0
 LD:'WHAT IS THE NAME OF THE'
  'ARRAY OF VARIABLE NAMES?'
  N←⎕
∇
```

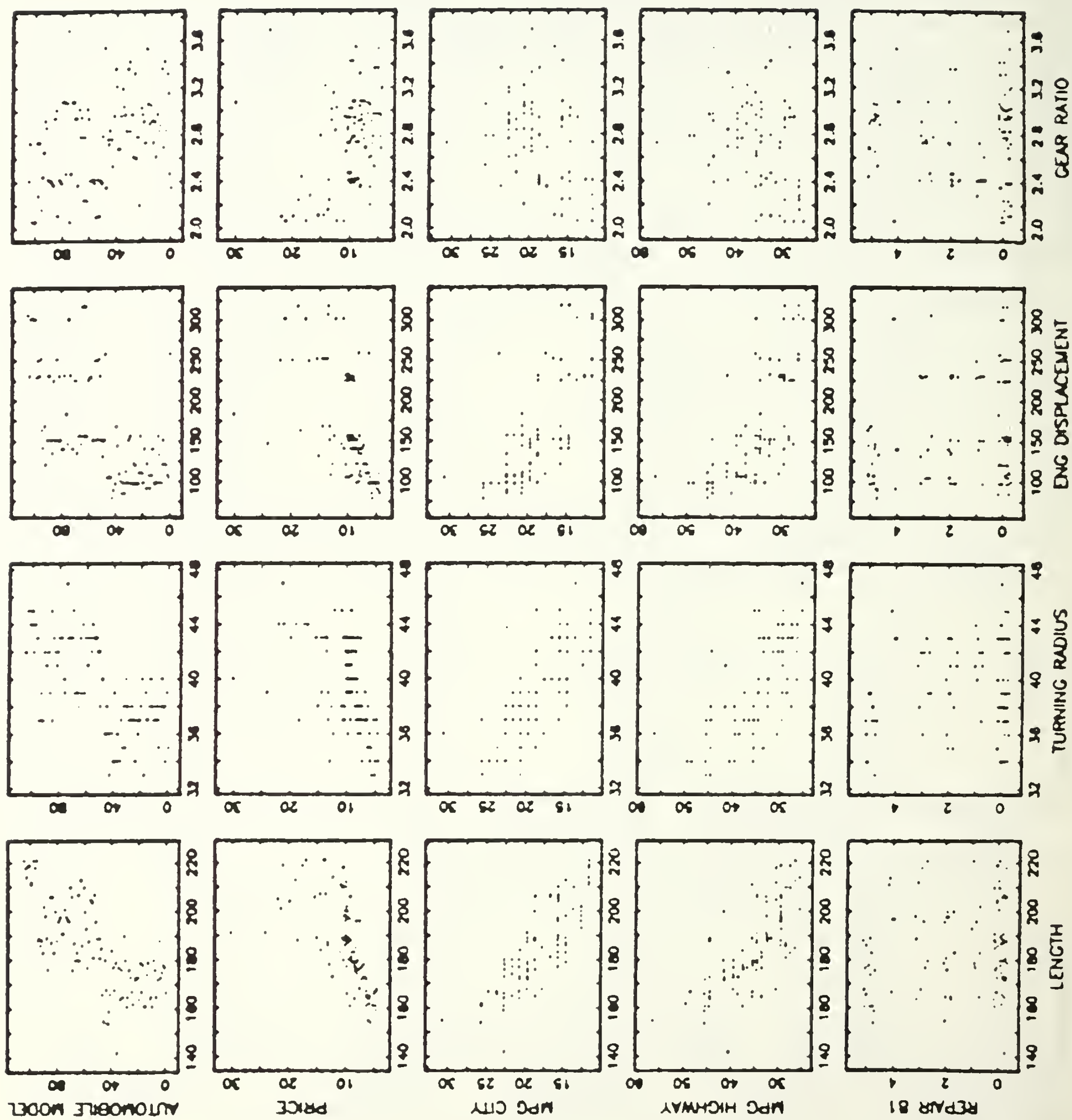
```
∇ SUB MATRIX;VR;VC;CI;RPOSN;CPOSN;CS;RZ1
  'ENTER AS A VECTOR THE VARIABLES'
  '(COLUMNS) FROM YOUR DATA SET WHICH'
  'YOU DESIRE TO BE DISPLAYED'
  CI←⎕
  'DO YOU DESIRE A SUBPOPULATION GROUP'
  'OUT OF ANY ONE VARIABLE?'
  'ENTER (Y OR N)'
  RZ1←⎕
  →(RZ1='N')/LD1
  'WHAT VARIABLE (COLUMN) IN THE ORIGINAL'
  'DATA IS THE SUB-GROUP?'
  VC←⎕
  'ENTER AS A VECTOR THE VALUES OF THE'
  'SUBPOPULATION GROUP THAT YOU WANT'
  VR←⎕
  →LD2
 LD1:VC←CI[1]
  VR←MATRIX[;VC]
  →LD2
 LD2:RPOSN←(MATRIX[;VC])∊VR
  DATA←RPOSN/[1] MATRIX
  CS←⍳CS←¯1↑CS←⍴MATRIX
  CPOSN←CS∊CI
  DATA←CPOSN/DATA
  'THE SUBDATA DESIRED IS A GLOBAL'
  'VARIABLE CALLED DATA AND HAS A'
  'SHAPE OF ',⍕(⍴DATA)
∇
```

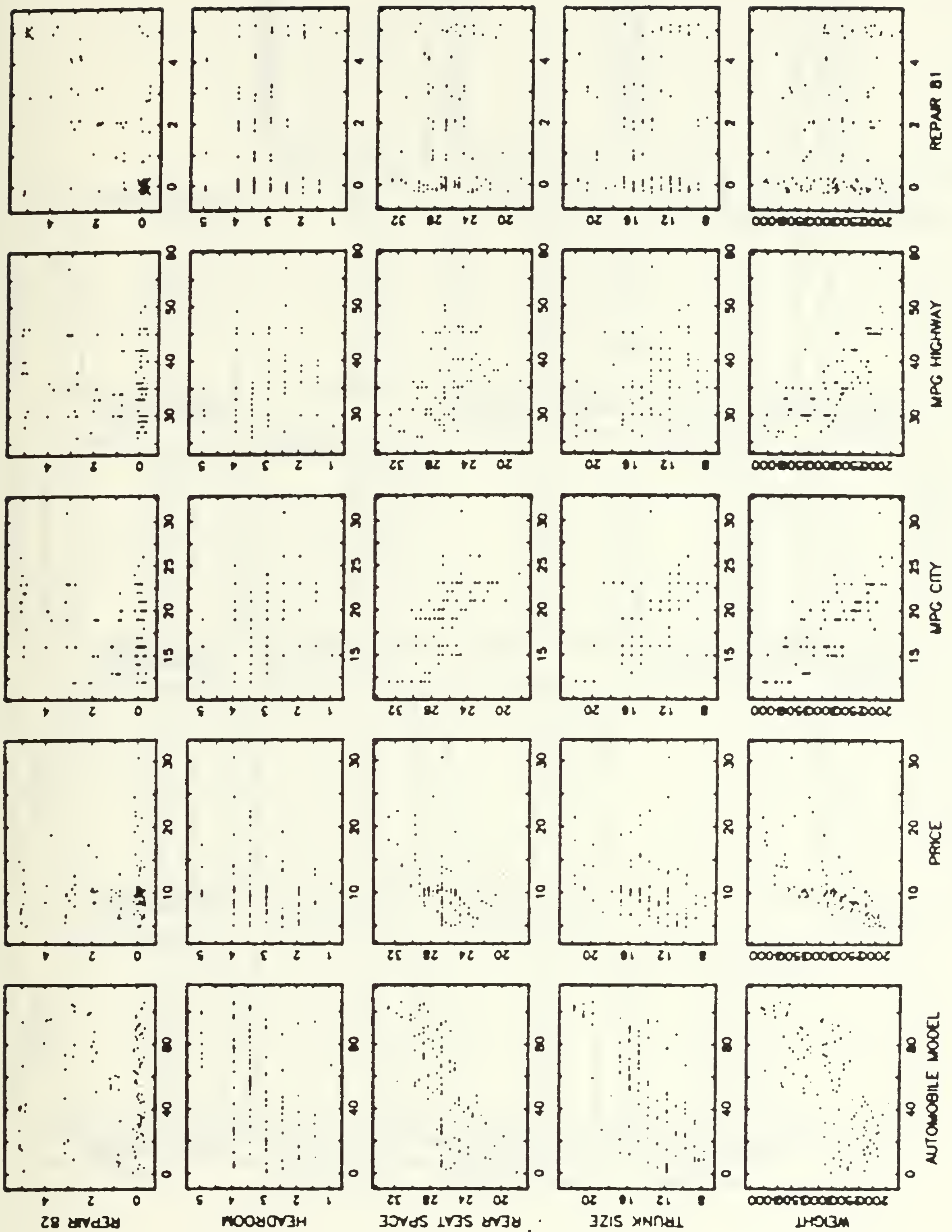
```
      ∇ LABELS DATA;IDX;I
[1]     IDX←¯1↑(⍴DATA)
[2]     'ENTER THE NAME OF EACH COLUMN '
[3]     'IN ORDER, THE NAME MUST CONSIST'
[4]     'OF NO MORE THAN 20 CHARACTERS'
[5]     'TO INCLUDE BLANK SPACES'
[6]     'NAME OF COLUMN 1 ?'
[7]     I←1
[8]     NAMES← 1 20 ⍴NAMES←20↑(⎕,'                    ')
[9]   LOOP:I←I+1
[10]    'NAME OF COLUMN ',(⍕I),'?'
[11]    NAMES←NAMES,[1](20↑(⎕,'                    '))
[12]    →(I<IDX)/LOOP
[13]    'THE COLUMN LABELS ARE A GLOBAL'
[14]    'VARIABLE CALLED NAMES'
      ∇


      ∇ JITTER;SIZE;TEMP;UI;RNGX;XMAX;XMIN;RES1;RES2;X;C;
          FRONT;REAR;PT;J
        X
[1]     'HOW MANY VARIABLES DO YOU DESIRE JITTERED?'
[2]     RES1←⎕
[3]     →(RES1=0)/0
[4]     C←0
[5]     'WHAT ARE THE VARS (COLUMNS) TO BE JITTERED?'
[6]     RES2←⎕
[7]   LOOPJ:C←C+1
[8]     (RES1=1)/PT←RES2
[9]     →(RES1=1)/JUMP
[10]    PT←RES2[C]
[11]  JUMP:X←DATA[;PT]
[12]    SIZE←(⍴X)-1
[13]    TEMP←(2÷SIZE)×((0,⍳SIZE)-(SIZE×0.5))
[14]    UI←TEMP[(⍴TEMP)?⍴TEMP]
[15]    RNGX←(XMAX←⌈/X)-(XMIN←⌊/X)
[16]    JX←0.05×RNGX×UI
[17]    X←X+JX
[18]    FRONT←DATA[;(⍳(PT-1))]
[19]    REAR←DATA[;((⍳(NCOL-PT))+PT)]
[20]    DATA←(FRONT,[2] X),[2] REAR
[21]    →(C<RES1)/LOOPJ


      ∇ TRANSFORM;RES;C;I;X;RES2;REAR;FRONT;A
[1]     'HOW MANY VARIABLES TO YOU WANT TO HAVE TR
[2]     RES←⎕
[3]     →(RES=0)/0
[4]     C←0
[5]     'WHAT ARE THE VARS (COLUMNS) TO BE TRANSF
[6]     RES2←⎕
[7]   LOOPA:C←C+1
[8]     (RES=1)/I←RES2
[9]     →(RES=1)/JUMP
[10]    I←RES2[C]
[11]  JUMP:X←DATA[;I]
[12]    'USING X HAS THE VAR, INPUT THE APL EXPRESSION'
[13]    'FOR THE TRANSFORMATION DESIRED ON COLUMN  ',⍕I
[14]    A←⎕
[15]    FRONT←DATA[;(⍳(I-1))]
[16]    REAR←DATA[;((⍳(NCOL-I))+I)]
[17]    DATA←(FRONT,[2] A),[2] REAR
[18]    →(C<RES)/LOOPA
```

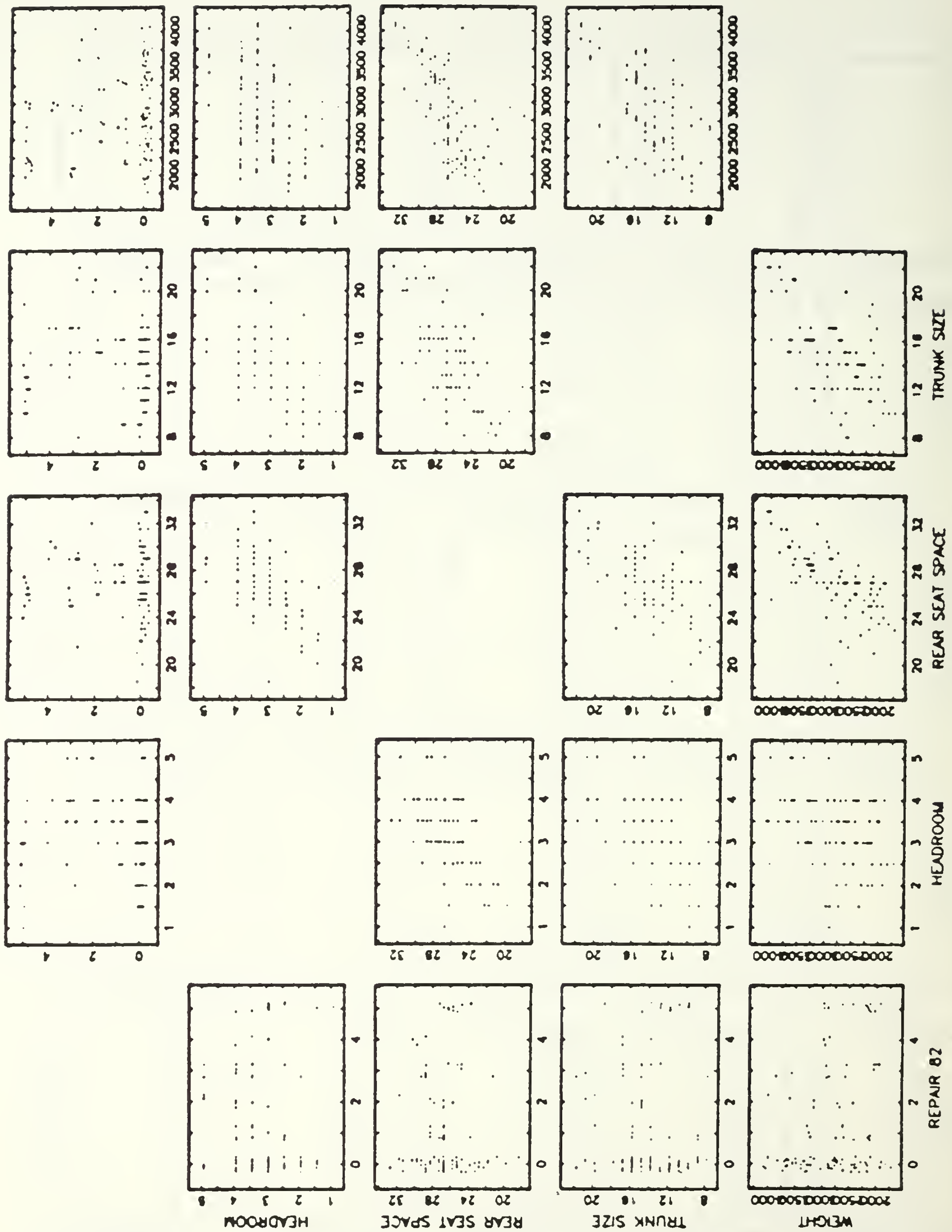
APPENDIX B; CAR DATA DISPLAYS

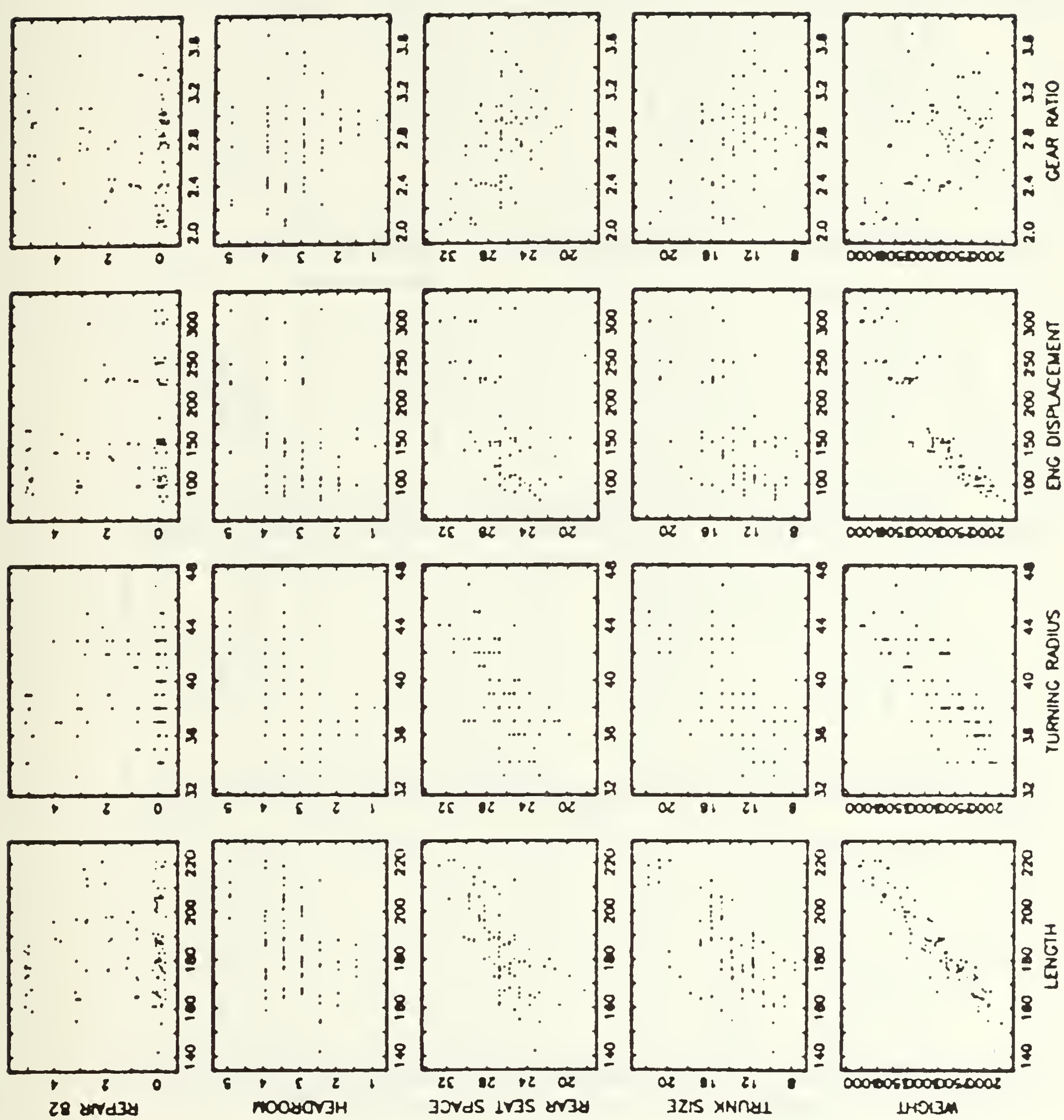


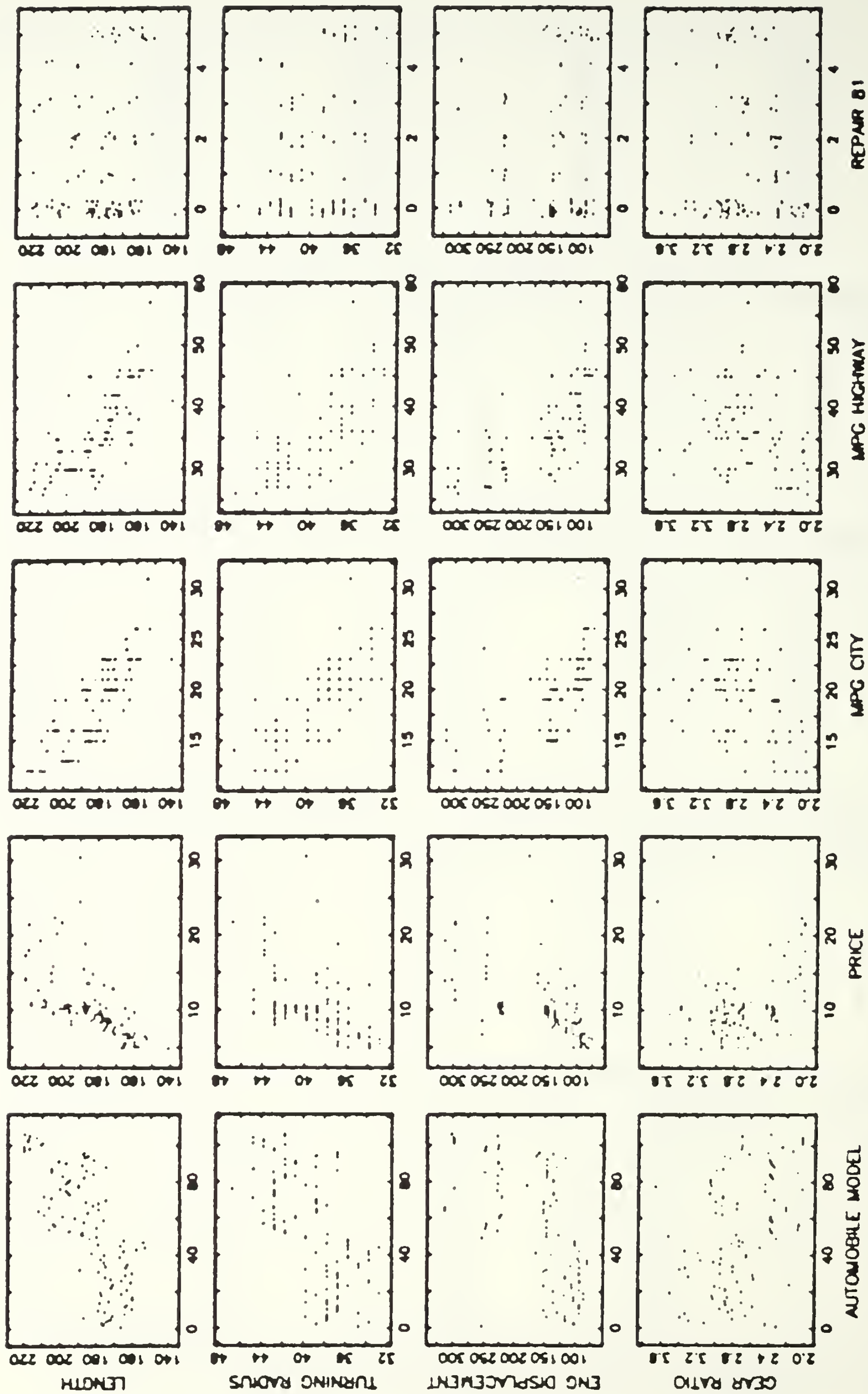

WEIGHT
TRUNK SIZE
REAR SEAT SPACE
HEADROOM
REPAIR 82
AUTOMOBILE MODEL
PRICE
MPG CITY
MPG HIGHWAY
REPAIR 81

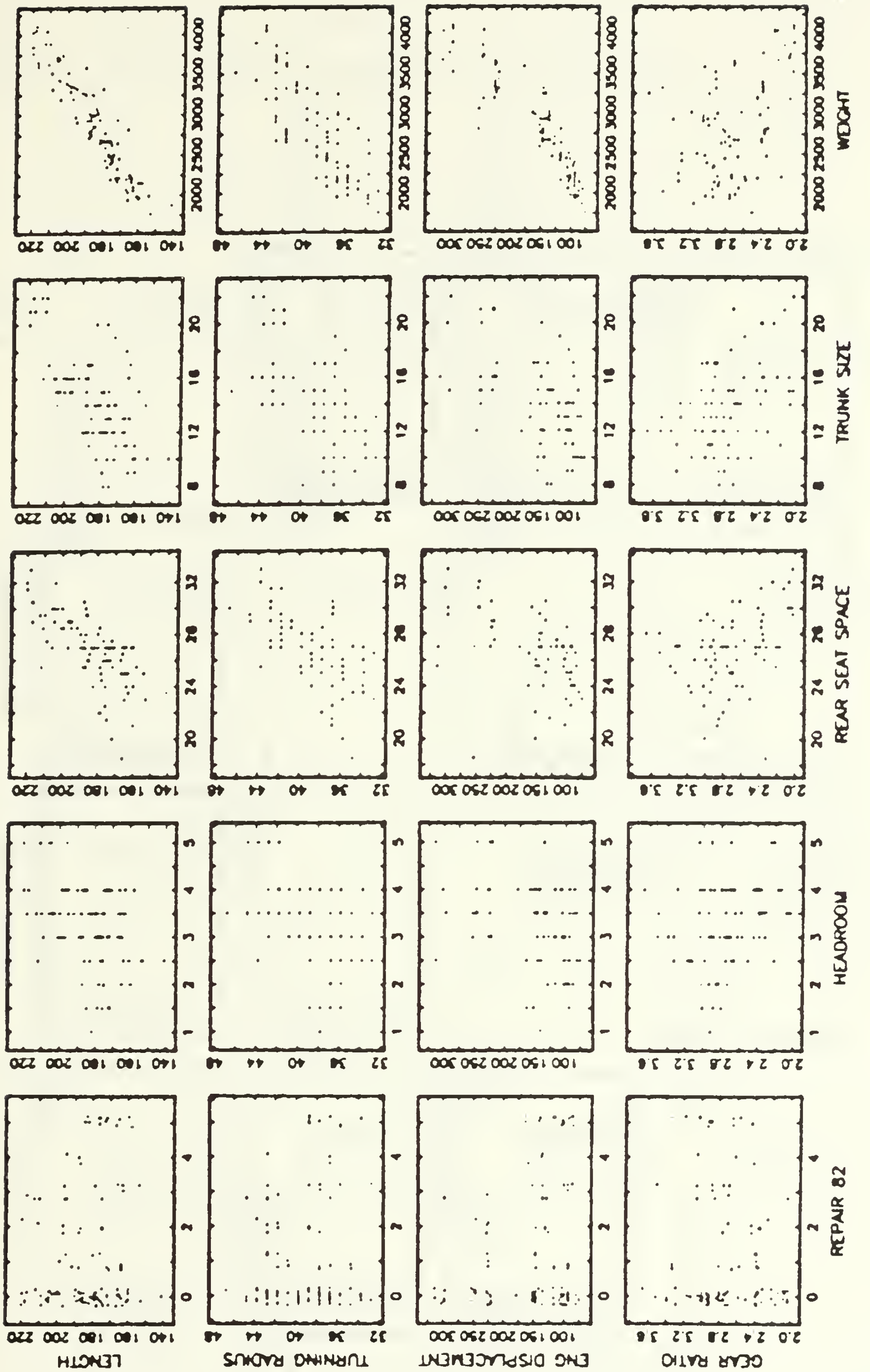

LENGTH
TURNING RADIUS
ENG DISPLACEMENT
GEAR RATIO
AUTOMOBILE MODEL
PRICE
MPG CITY
MPG HIGHWAY
REPAIR 81

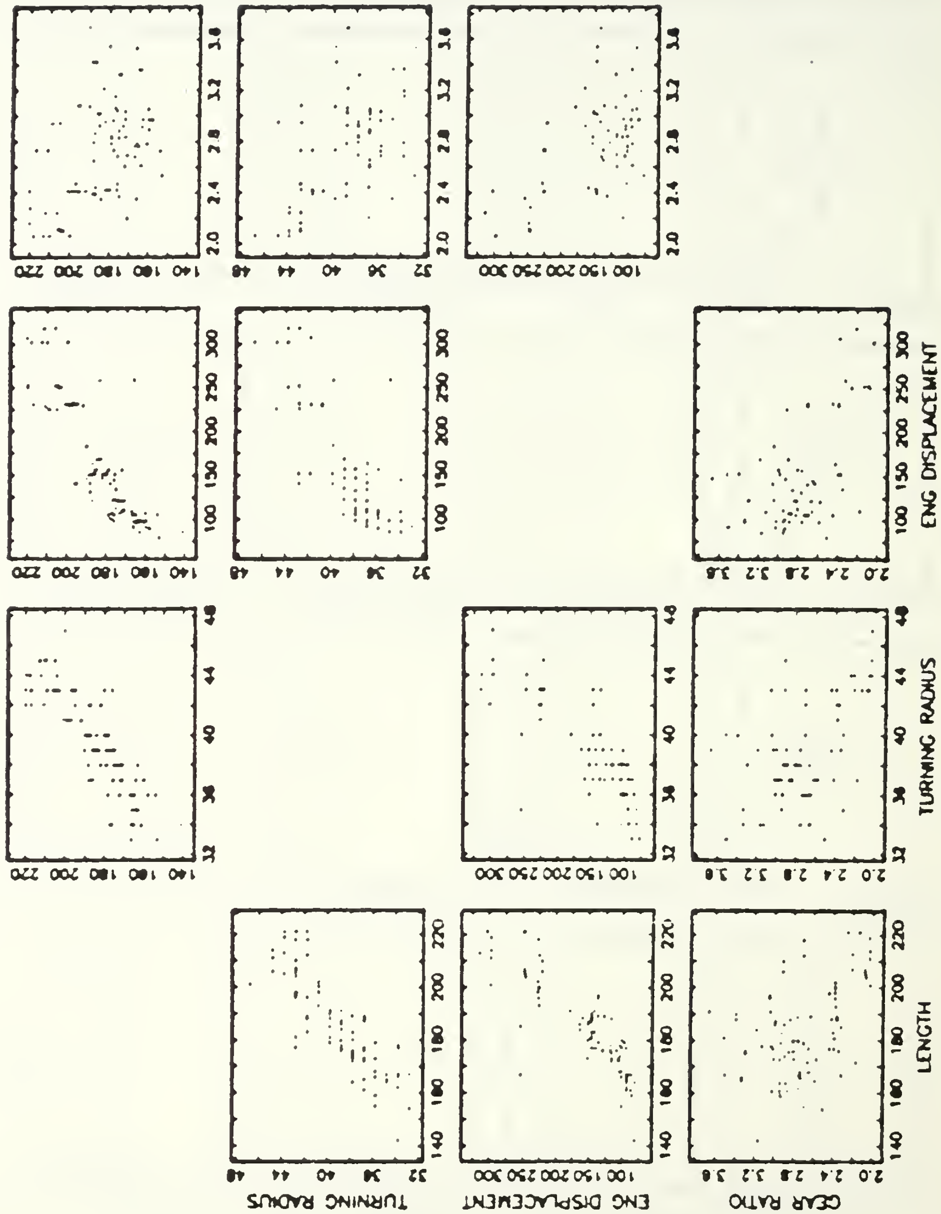

REPAIR 81
MPG HIGHWAY
MPG CITY
PRICE
AUTOMOBILE MODEL
REPAIR 82
HEADROOM
REAR SEAT SPACE
TRUNK SIZE
WEIGHT


WEIGHT
TRUNK SIZE
REAR SEAT SPACE
HEADROOM
REPAIR 82

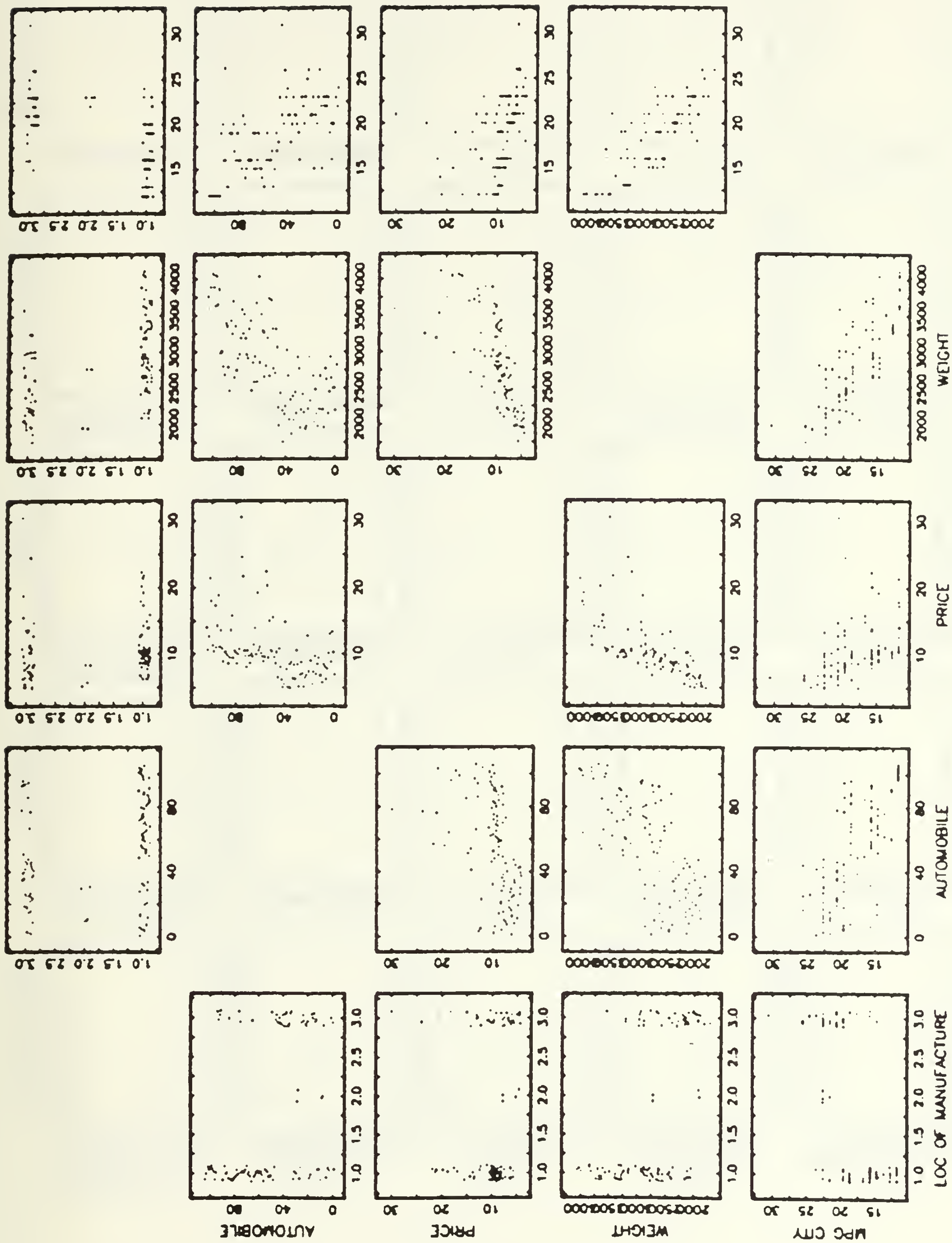

GEAR RATIO
ENG DISPLACEMENT
TURNING RADIUS
LENGTH
REPAIR 82
HEADROOM
REAR SEAT SPACE
TRUNK SIZE
WEIGHT

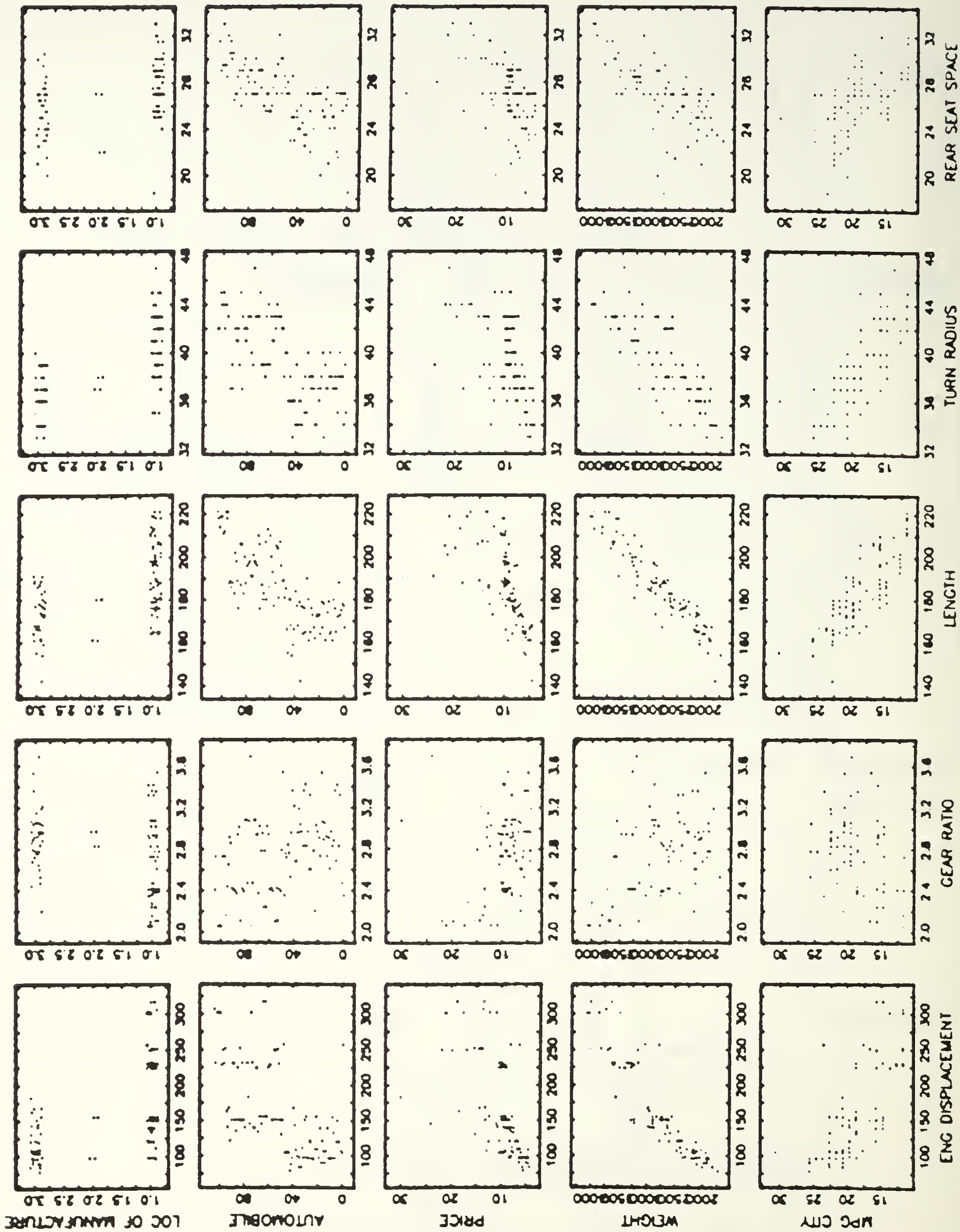

LENGTH
TURNING RADIUS
ENG DISPLACEMENT
GEAR RATIO
AUTOMOBILE MODEL
PRICE
MPG CITY
MPG HIGHWAY
REPAIR 81

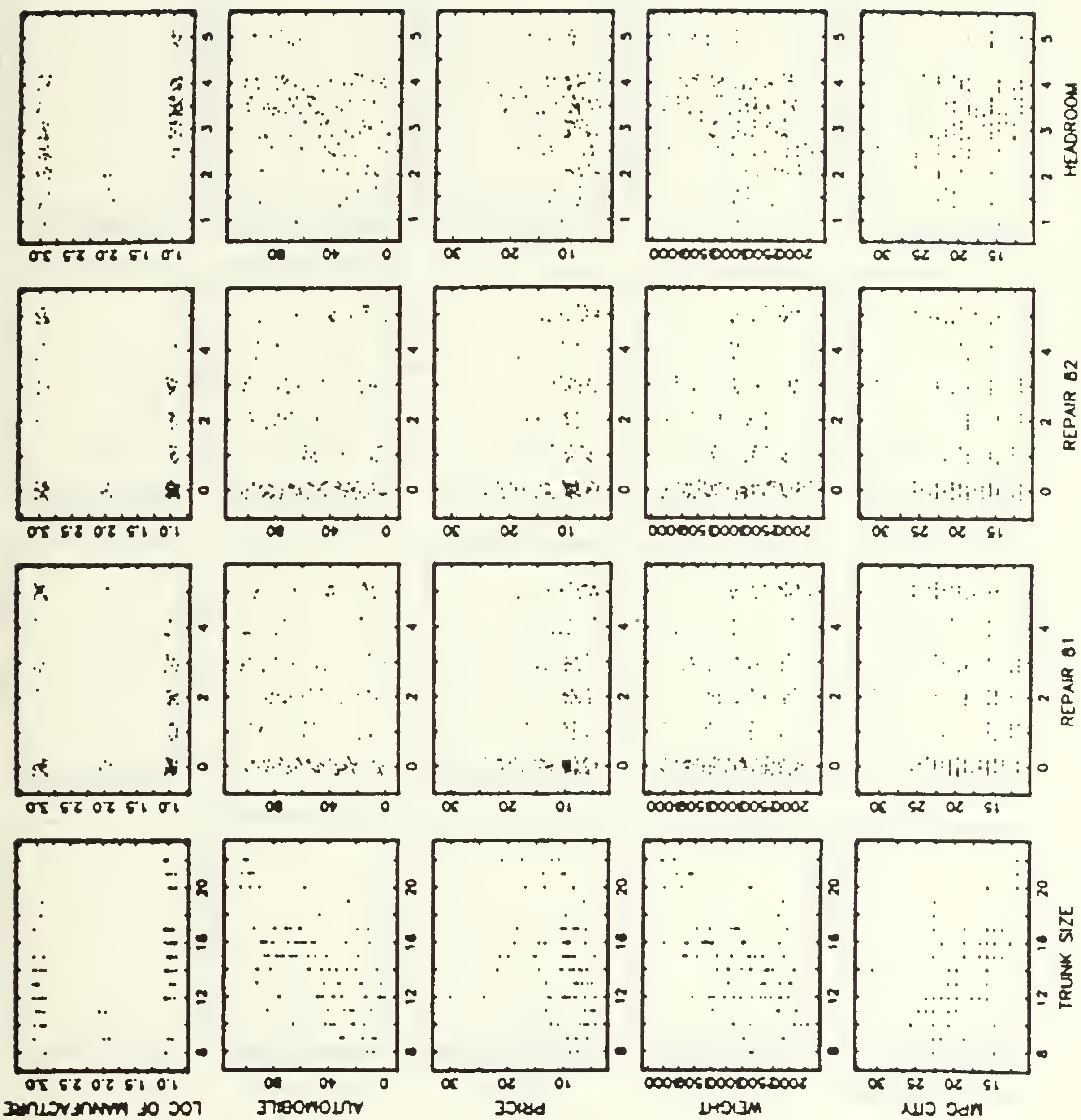

WEIGHT
TRUNK SIZE
REAR SEAT SPACE
HEADROOM
REPAIR 82
LENGTH
TURNING RADIUS
ENG DISPLACEMENT
GEAR RATIO

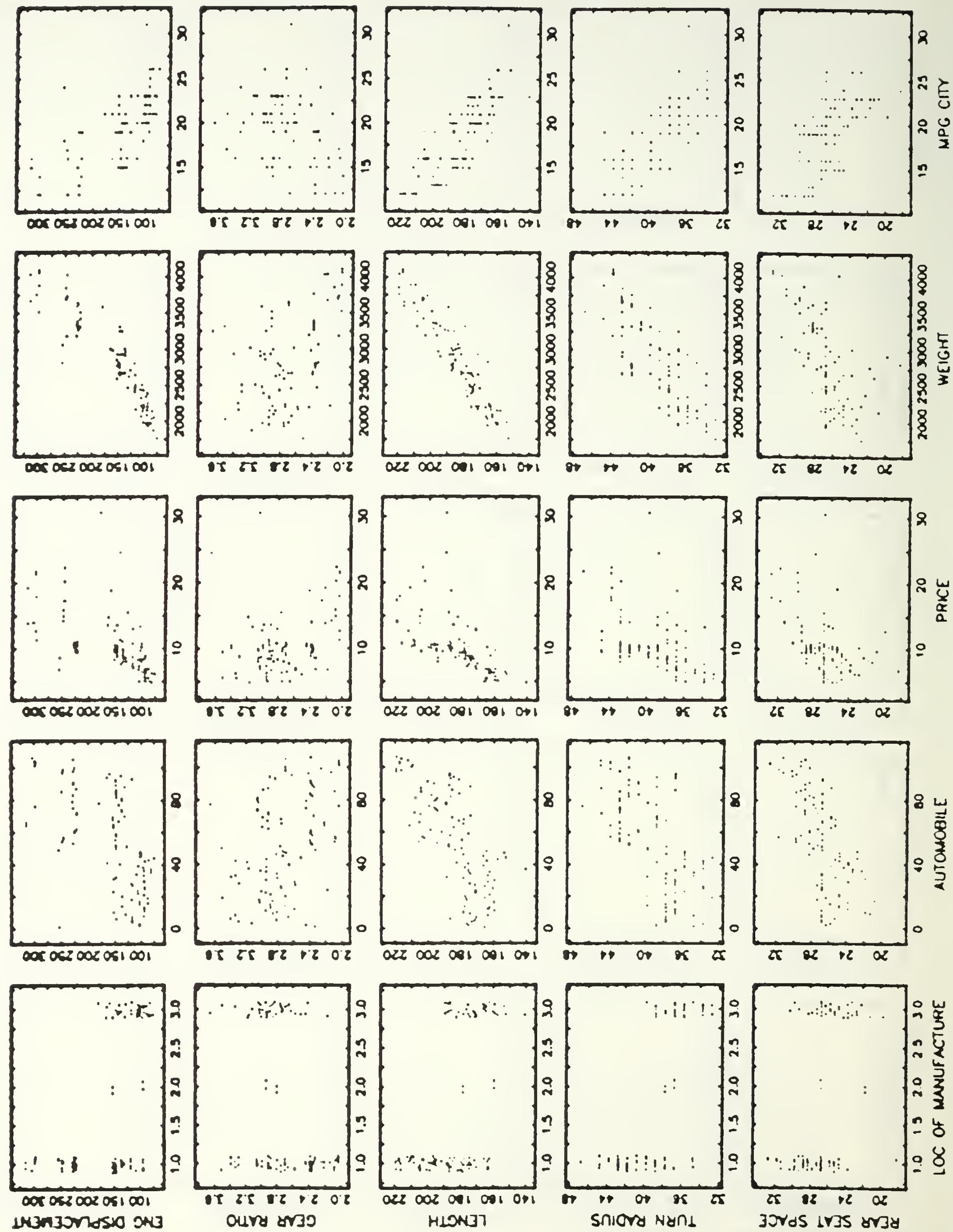

LENGTH
TURNING RADIUS
ENG DISPLACEMENT
GEAR RATIO

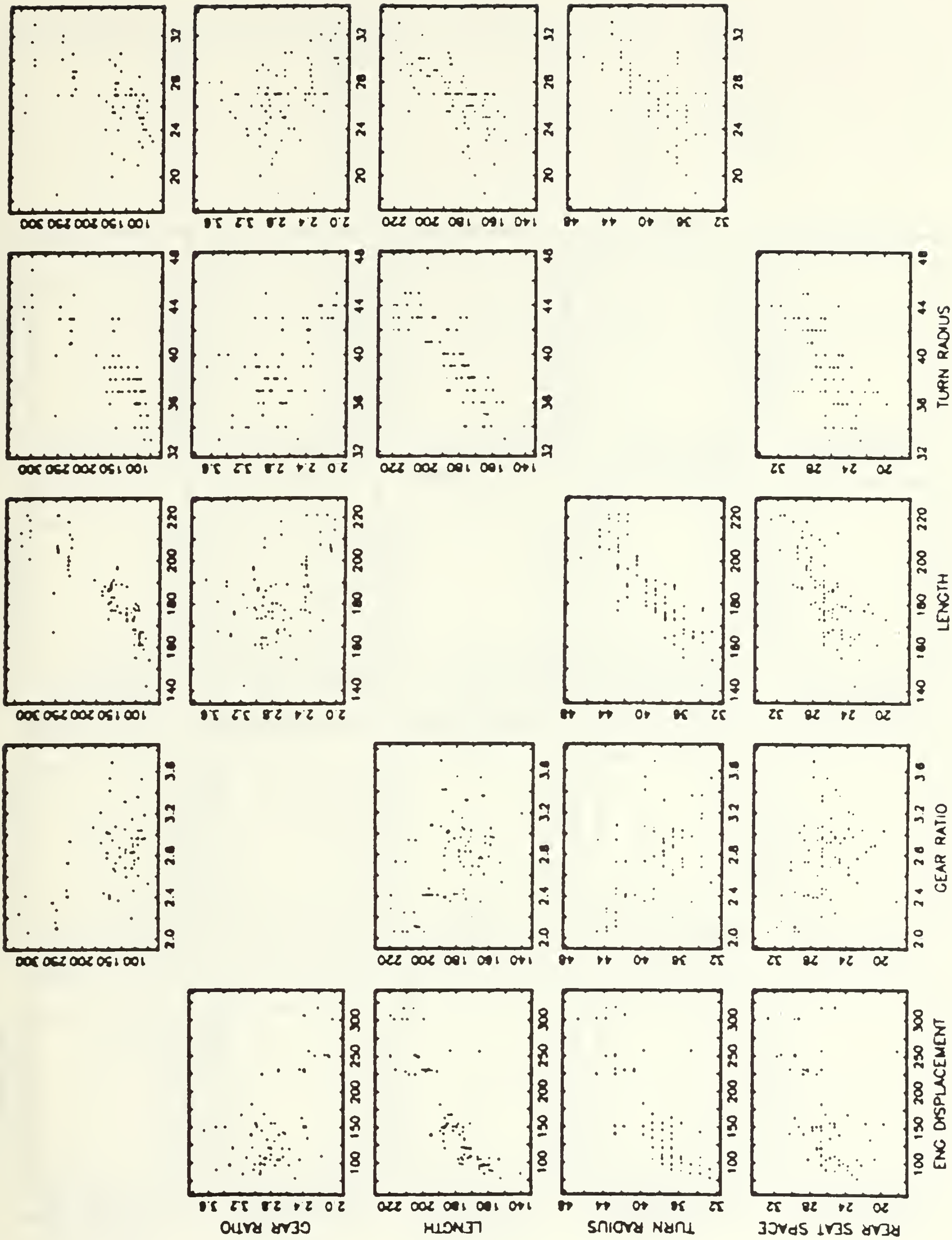
B. ENHANCED DISPLAY

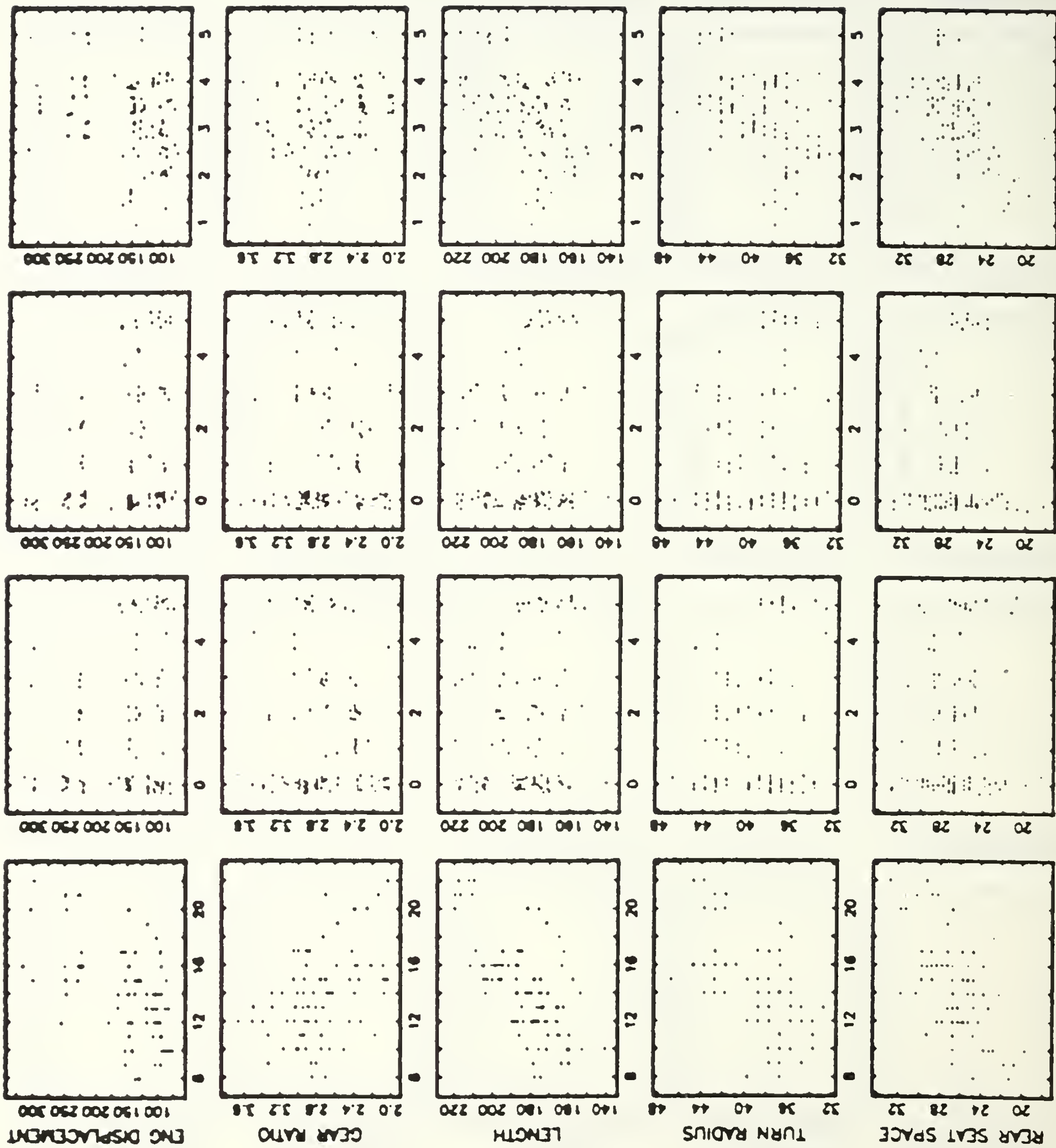


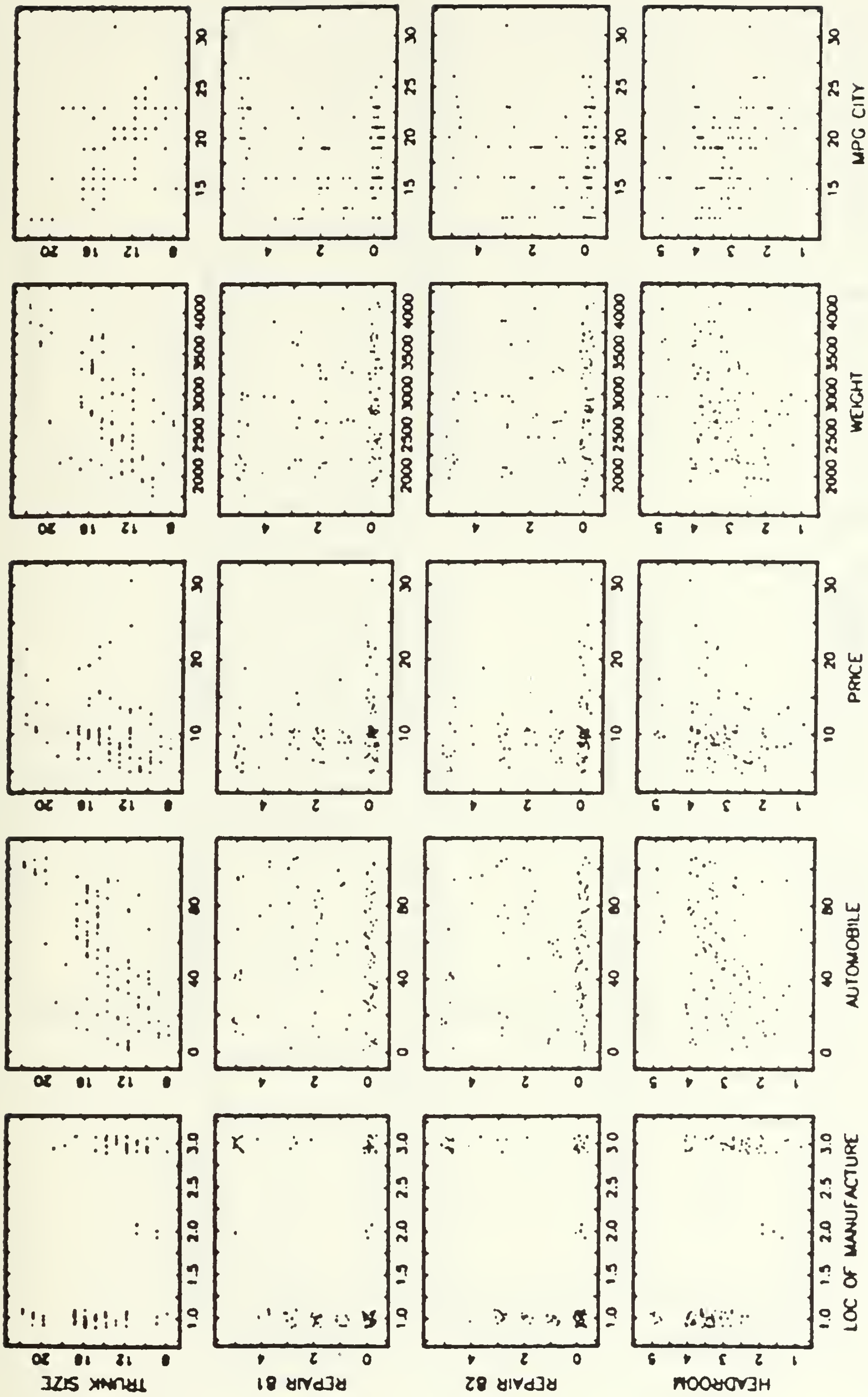

LOC OF MANUFACTURE
AUTOMOBILE
PRICE
WEIGHT
MPG CITY
ENG DISPLACEMENT
GEAR RATIO
LENGTH
TURN RADIUS
REAR SEAT SPACE

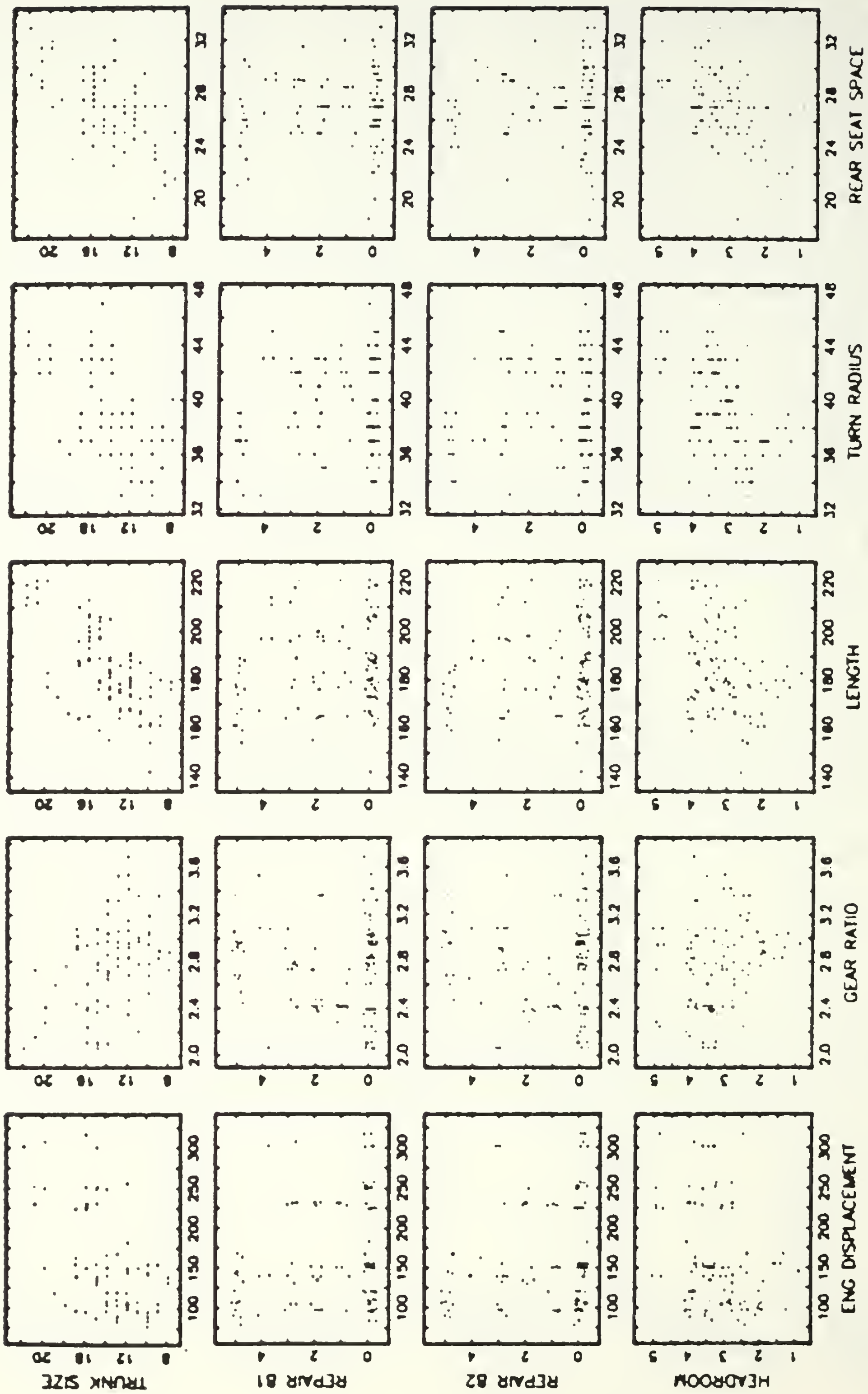

HEADROOM
REPAIR 82
REPAIR 81
TRUNK SIZE
LOC OF MANUFACTURE
AUTOMOBILE
PRICE
WEIGHT
MPG CITY

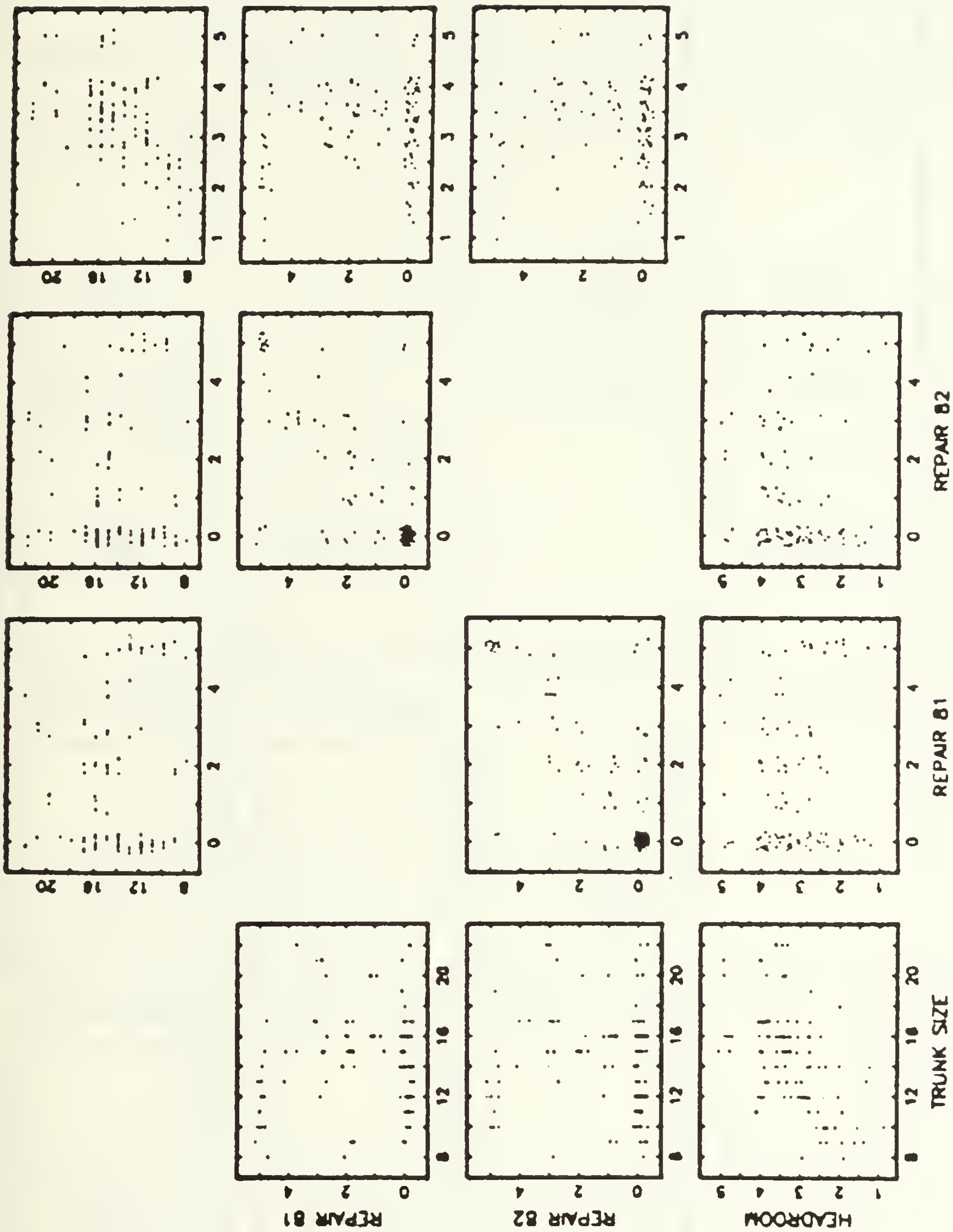

MPG CITY
WEIGHT
PRICE
AUTOMOBILE
LOC OF MANUFACTURE
ENG DISPLACEMENT
GEAR RATIO
LENGTH
TURN RADIUS
REAR SEAT SPACE


ENG DISPLACEMENT
GEAR RATIO
LENGTH
TURN RADIUS
REAR SEAT SPACE

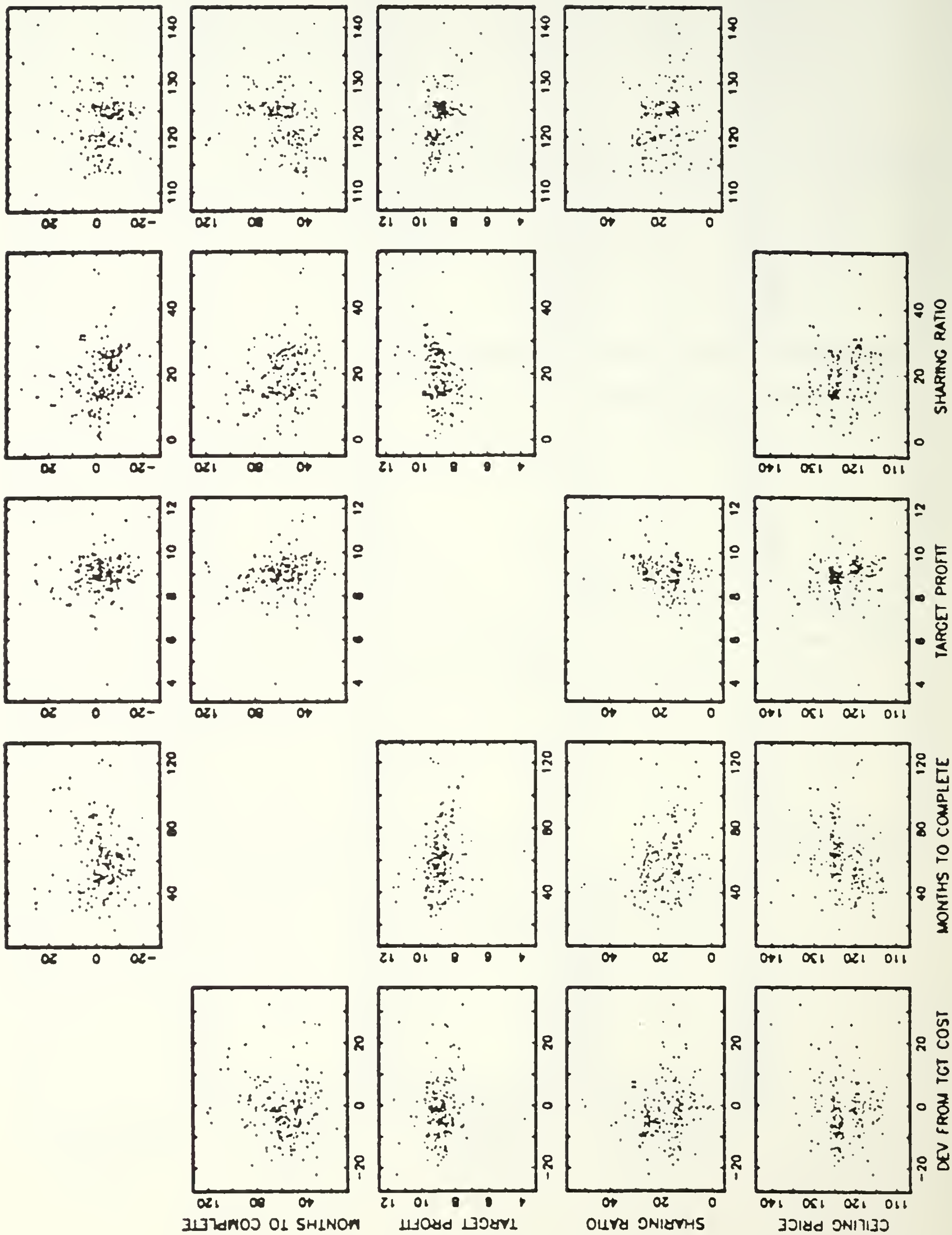

HEADROOM
REPAIR 82
REPAIR 81
TRUNK SIZE
ENG DISPLACEMENT
GEAR RATIO
LENGTH
TURN RADIUS
REAR SEAT SPACE


MPG CITY
WEIGHT
PRICE
AUTOMOBILE
LOC OF MANUFACTURE
TRUNK SIZE
REPAIR 81
REPAIR 82
HEADROOM

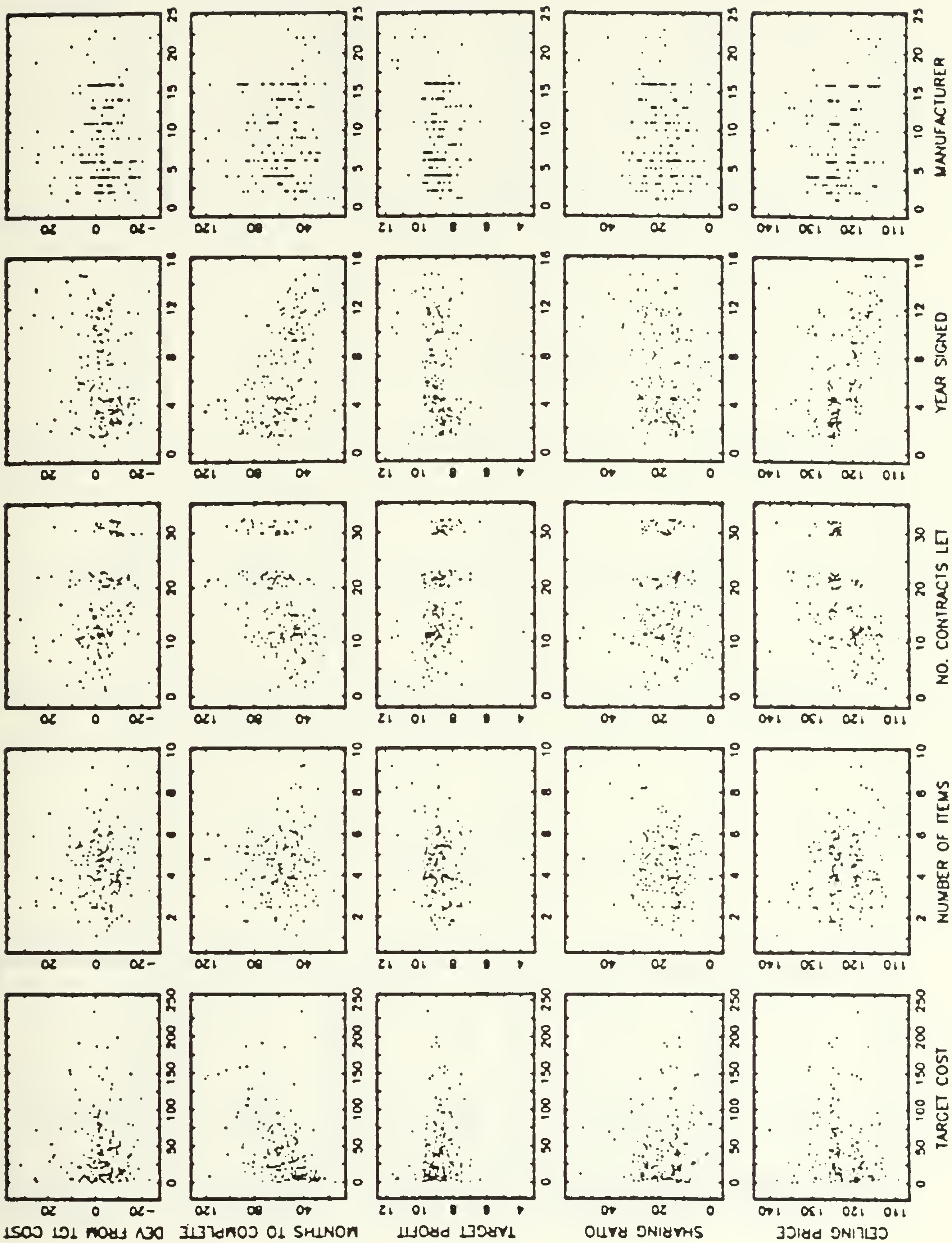

TRUNK SIZE
REPAIR B1
REPAIR B2
HEADROOM
ENG DISPLACEMENT
GEAR RATIO
LENGTH
TURN RADIUS
REAR SEAT SPACE


TRUNK SIZE
REPAIR 81
REPAIR 82
HEADROOM

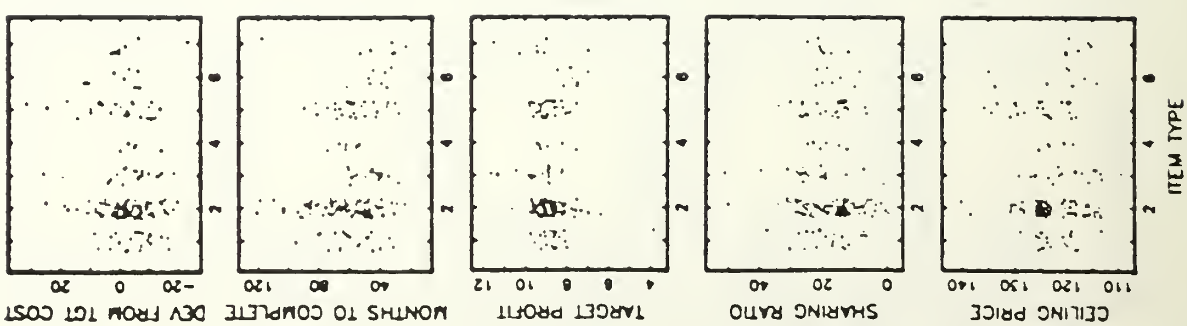
APPENDIX C



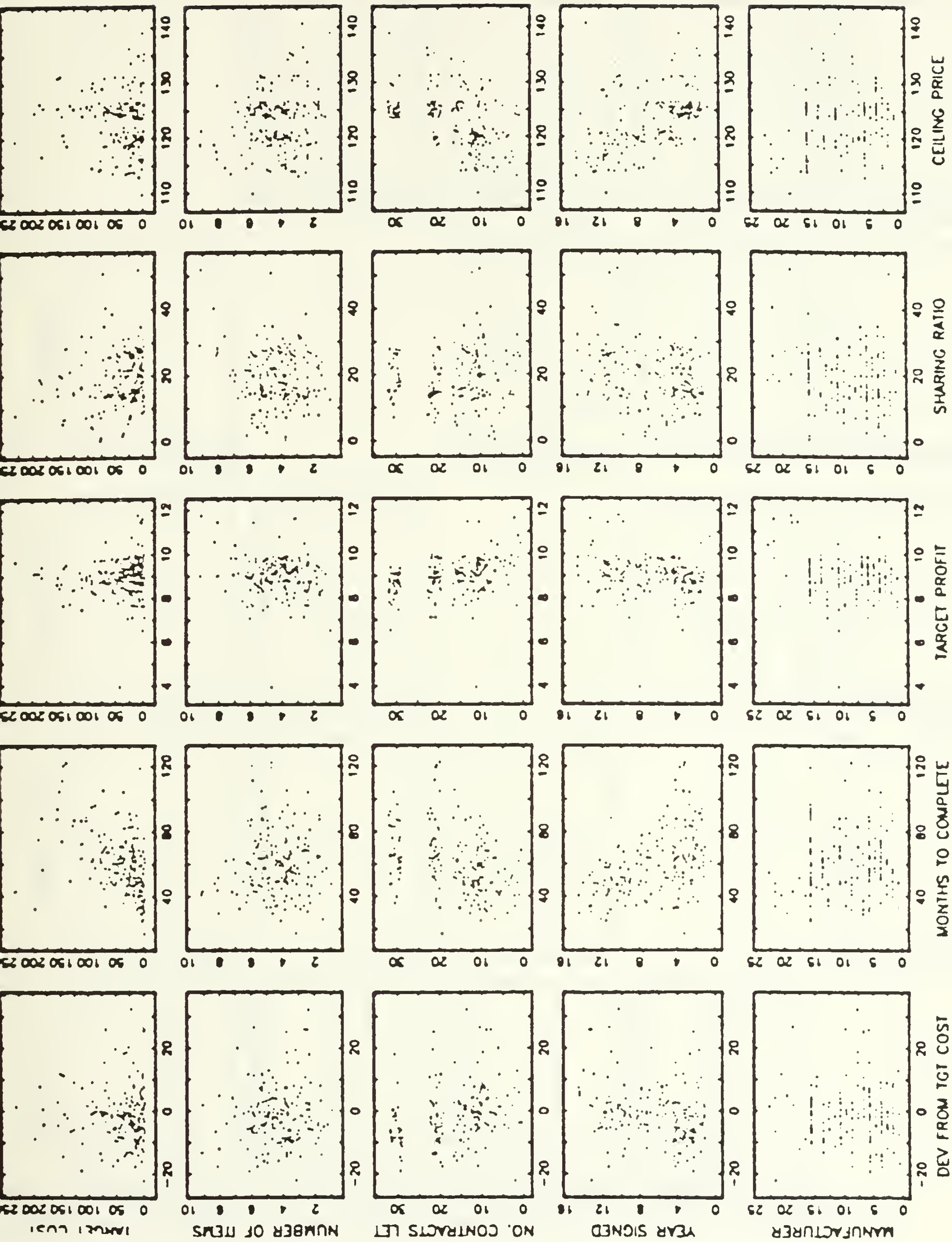
SHARING RATIO
TARGET PROFIT
MONTHS TO COMPLETE
DEV FROM TGT COST
CEILING PRICE

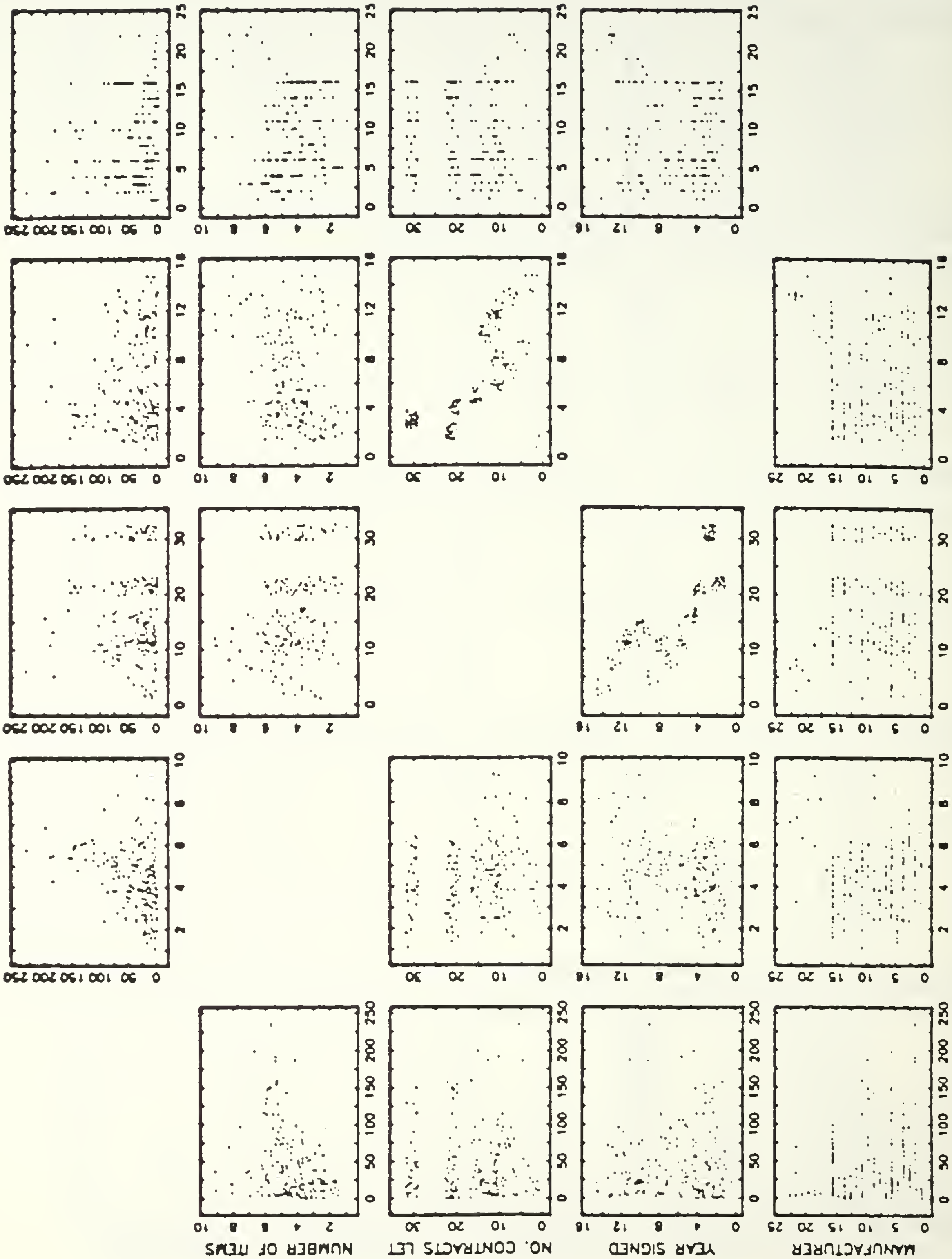


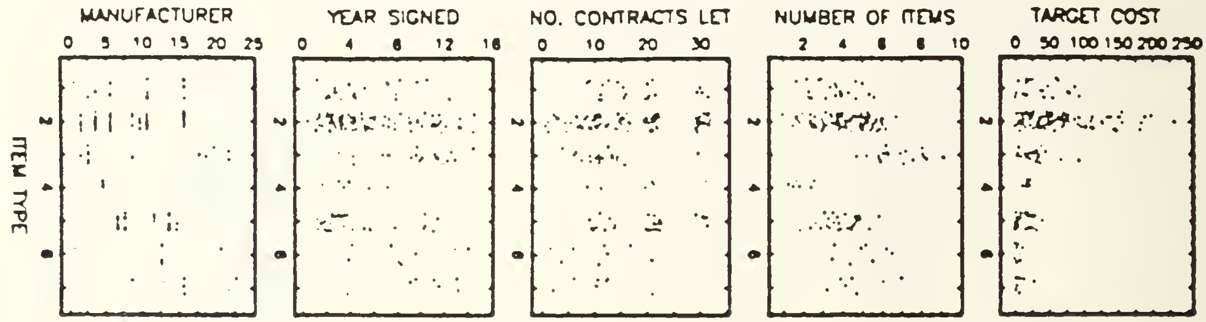
MANUFACTURER
YEAR SIGNED
NO. CONTRACTS LET
NUMBER OF ITEMS
TARGET COST
DEV FROM TGT COST
MONTHS TO COMPLETE
TARGET PROFIT
SHARING RATIO
CEILING PRICE

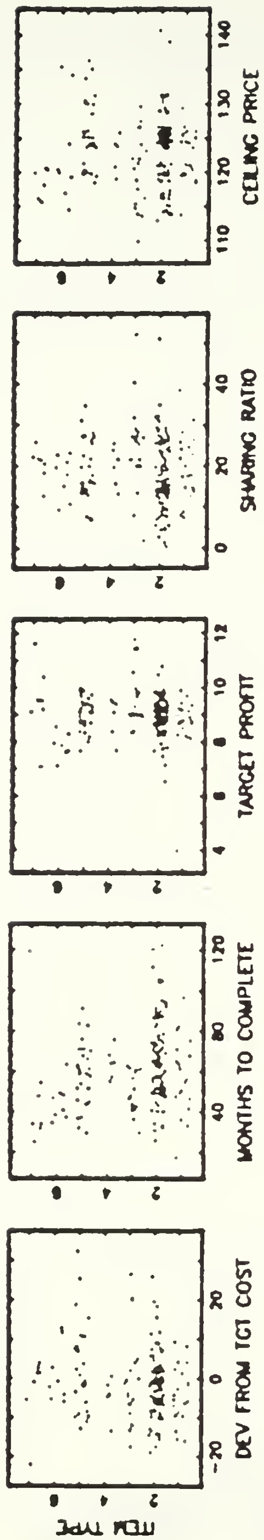


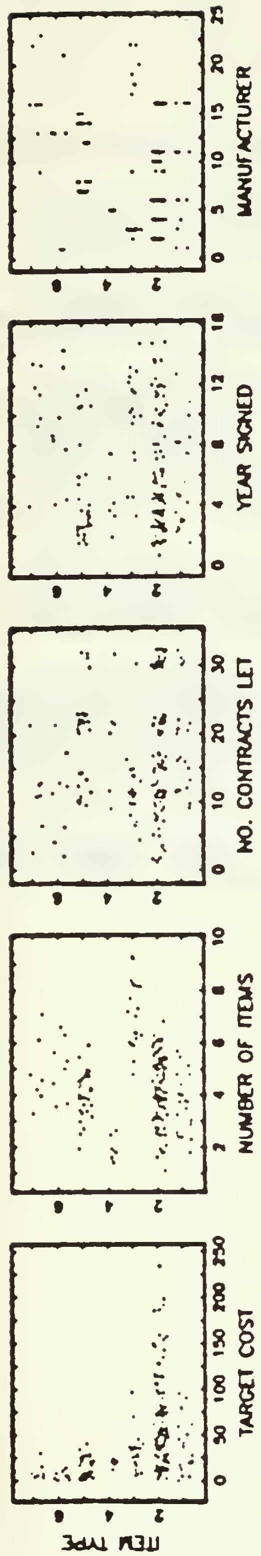
CEILING PRICE
SHARING RATIO
TARGET PROFIT
MONTHS TO COMPLETE
DEV FROM TGT COST
ITEM TYPE


CEILING PRICE
SHARING RATIO
TARGET PROFIT
MONTHS TO COMPLETE
DEV FROM TGT COST
NUMBER OF ITEMS
NO. CONTRACTS LET
YEAR SIGNED
MANUFACTURER


TARGET COST
NUMBER OF ITEMS
NO. CONTRACTS LET
YEAR SIGNED
MANUFACTURER


MANUFACTURER
YEAR SIGNED
NO. CONTRACTS LET
NUMBER OF ITEMS
TARGET COST
ITEM TYPE


ITEM TYPE
DEV FROM TGT COST
MONTHS TO COMPLETE
TARGET PROFIT
SHARING RATIO
CEILING PRICE


ITEM TYPE
TARGET COST
NUMBER OF ITEMS
NO. CONTRACTS LET
YEAR SIGNED
MANUFACTURER

LIST OF REFERENCES

1. Chambers, J.M., and others. Graphical Methods for Data Analysis, Wadsworth, 1983.

2. Consumer Union, 1984 Buying Guide, December 1983.

3. Hoaglin, D.C., and others, Understanding Robust and Exploratory Data Analysis, Wiley, 1983.

4. "GRAFSTAT on the IBM 3277 Display", Computer Center Newsletter, Naval Postgraduate School, 18 May 1983.

5. Mar, D., VS APL at NPS, TN-VM-10, Naval Postgraduate School, July 1982.

6. Reed, R.R., A Primer in APL, Naval Postgraduate School, Department of Operations Research, 1983, unpublished.

7. Dixon, M.W., A Statistical Analysis of Deviations from Target Cost in NAVAIRSYSCOMHQ Fixed-Price Incentive Contracts, Masters Thesis, Naval Postgraduate School, March 1973.

8. Rand Corporation Report RM-5120-PR, Cost Incentives and Contract Outcomes: An Emperical Analysis, by Irving N. Fisher, September 1966.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, code 0142 Naval Postgraduate School Monterey, California 93943	2
3.	Professor Peter W. Lewis Code 55Lw Department of Operations Research Naval Postgraduate School Monterey, California 93943	24
4.	Professor John Orav Code 55 Department of Operations Research Naval Postgraduate School Monterey, California 93843	1
5.	Captain D.A. Thomas 1807 Catawba Street Fayetville, North Carolina 28303	1
6.	Captain M.D. Johnson 581A Wilkes Lane Monterey, California 93940	2

210293

Thesis
J6237
c.1

Johnson
Draftsman displays, a graphical technique for exploratory data analysis.